DOCUMENT RESUME

ED 360 315 TM 019 340

AUTHOR	Afflerbach, Peter, Ed.
TITLE	Issues in Statewide Reading Assessment.
INSTITUTION	ERIC Clearinghouse on Tests, Measurement, and Evaluation, Washington, DC.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO	ISBN-0-89785-216-X
PUB DATE	Dec 90
CONTRACT	RI88062003
NOTE	160p.
PUB TYPE	Collected Works - General (020) -- Information Analyses - ERIC Clearinghouse Products (071)
EDRS PRICE	MF01/PC07 Plus Postage.
DESCRIPTORS	Decision Making; *Educational Assessment; Educational Practices; Elementary Secondary Education; Evaluation Utilization; Literacy; *National Surveys; *Reading Achievement; Reading Instruction; Reading Tests; *State Programs; Student Evaluation; Testing Programs; *Test Use; Test Validity
IDENTIFIERS	Alternative Assessment

ABSTRACT

This paper presents six chapters that describe how statewide reading assessment is currently being performed and how the data are being used. The validity of statewide reading assessment instruments and the appropriate uses of statewide reading assessment data are explored. Several chapters discuss new ways in which some states conduct reading assessment, while others suggest alternative and complementary forms of reading assessment. The range of issues is intended to help in the assessment of relative strengths and weaknesses of current statewide reading practice and consider future directions in reading assessment. The following six chapters are provided: (1) "The Call for Assessment of Reading at the Statewide Level" (Peter Afflerbach); (2) "Developing a Statewide Reading Assessment Program" (Linda Hansche); (3) "Issues in Early Childhood Assessment" (William H. Teale); (4) "The Role of Teacher-Based Information in Statewide Assessments of Literacy Learning" (Elfrieda H. Hiebert); (5) "National Survey of the Use of Test Data for Educational Decision Making" (Sheila W. Valencia); and (6) "Statewide Reading Assessment: A Survey of the States" (Peter Afflerbach). Charts for each of the 50 states are included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 360 315

ISSUES IN STATEWIDE READING ASSESSMENT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

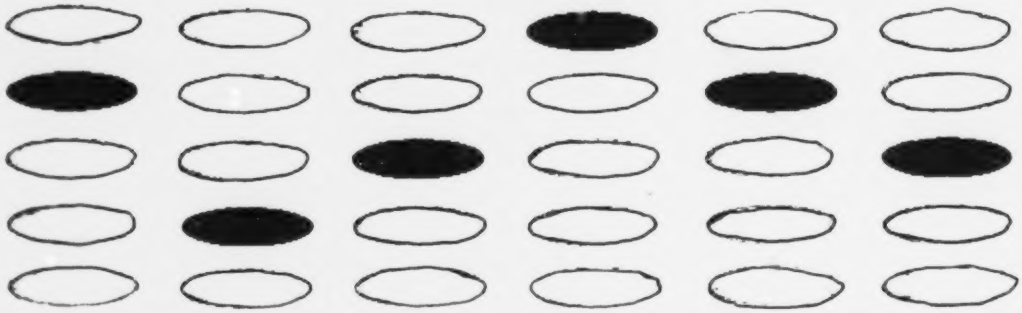
☑ This document has been reproduced as received from the person or organization originating it

☐ Minor changes have been made to improve reproduction quality

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

Peter Afflerbach, Editor
University of Maryland, College Park

Published by:
The ERIC Clearinghouse on Tests, Measurement, and Evaluation



American Institutes for Research
3333 K Street, NW, Washington, DC 20007

TM 019 340

BEST COPY AVAILABLE

Published by:
The ERIC Clearinghouse on Tests, Measurement, and Evaluation
American Institutes for Research
3333 K Street, NW
Washington, DC 20007

Printed in the United States of America

This publication was prepared, in part, with funding from the Office of Educational Research and Improvement (OERI), U.S. Department of Education, under contract R-88-062003. The opinions expressed in this report do not necessarily reflect the positions or policy of OERI or the Department of Education.

Library of Congress Cataloging in Publication Data

What's happening in statewide reading assessment / Peter Afflerbach, editor.

ISBN 0-89785-216-X
1. Reading--United States--Ability testing. I. Afflerbach, Peter
LB1050.46.W48 1990
428.4'076--dc20 90-34695
CIP

Issues in Statewide Reading Assessment

ISBN 0-89785-216-X

First Printing December 1990

Contents

Introduction 1

The Call for Assessment of Reading at the Statewide Level 7

Developing a Statewide Reading Assessment Program 17

Issues in Early Childhood Assessment 35

The Role of Teacher-Based Information in Statewide Assessments of Literacy Learning 57

National Survey of the Use of Test Data for Educational Decision Making 75

Statewide Reading Assessment: A Survey of the States 101

Index 163

Introduction

What's the purpose of this volume?

This volume describes how statewide reading assessment currently is being performed and how the data are being used. We explore the validity of statewide reading assessment instruments and the appropriate uses of statewide reading assessment data. Several contributors discuss new ways that some states are conducting reading assessment. Other contributors suggest alternative and complementary forms of reading assessment. We hope that the range of issues considered by the authors of this volume will help you assess the relative strengths and weaknesses of statewide reading assessment practice and consider future directions in reading assessment.

What's in the volume?

The Call for Reading Assessment at the Statewide Level

In this chapter, Peter Afflerbach of the University of Maryland describes some of the reasons for the current popularity of statewide reading assessments. These include the call for schools' accountability; the belief that current assessment instruments provide objective, scientific, and valid data on students' reading ability; and the traditional link between assessment and the maintenance of educational standards.

Afflerbach discusses some concerns about statewide reading assessments. These concerns include the validity of statewide reading assessment instruments and the disparity between the development in our understanding of the reading process and how the reading process is assessed. He also discusses the potential effects of statewide reading assessment on reading curriculum and the teaching of reading are also discussed.

Developing a Statewide Reading Assessment Program

Much of this volume concentrates on the products of statewide reading assessment. This chapter describes the process of how one state developed a statewide reading assessment program. In the chapter, Linda Hansche, Director of the Georgia Assessment Project, describes the development, implementation, and ongoing revision of one state's reading assessment program.

Hansche follows the development of a criterion-referenced reading test from its inception as a legislative mandate. She reviews the process of determining goals for the program, writing and reviewing items, field testing and reviewing bias, constructing operational assessment forms, and setting standards.

Hansche describes the contributions of different groups throughout the development of the reading assessment program. These groups include teachers, administrators, parents, legislators, and business people. Hansche's description of the development of a reading assessment program will familiarize you with the processes and sources of input involved in the development of statewide reading assessment.

Issues in Early Childhood Reading Assessment

In this chapter, William Teale of the University of Texas at San Antonio describes what is known about young children's literacy development and how children's literacy is assessed. While the understanding of children's emergent literacy has evolved, a critical gap still exists between what is known about young children's developing literacy and how it is assessed. The chapter describes the nature of current assessment of young children's literacy and concerns with many of the current approaches. These concerns include the developmentally inappropriate content and format of statewide assessments and how the assessments may influence early reading curriculum.

Next, Teale describes several programs which currently use alternative methods of assessment. These programs represent a re-thinking of issues related to early childhood reading assessment, and because they are based on our understanding of young children's language development, they may be more ecologically valid. Teale concludes that statewide assessment programs have the opportunity to be an important, positive influence on the development of reading assessments that will be sensitive to the nature of young children's literacy development.

The Role of Teacher-Based Information in Statewide Assessments of Literacy Learning

In this chapter, Elfrieda Hiebert of the University of Colorado at Boulder describes teacher-based measures of students' reading ability and stresses the importance of both reading curriculum and assessment, which are "inextricably interwoven." Hiebert proposes that teacher-based assessment allows for more accurate assessment of "higher literacies," and that acceptance and use of teacher-based assessment will encourage more teacher involvement in the educational enterprise. Hiebert also suggests that when teacher-based assessment is embedded in instruction, the link between assessment and instruction is strengthened.

Hiebert describes different types of teacher-based reading assessment, including the gathering of data through observing, questioning, and interviewing students, and the examination of students' work samples and portfolios. She also explores programs in the United States and abroad which successfully combine teacher-based assessment with statewide product-oriented assessments.

In conclusion, Hiebert considers several issues which must be addressed in order for such teacher-based assessments to become a valued and valuable part of statewide reading assessment. These include fostering change in peoples' conceptions of the forms that accurate assessment can take and changing teacher education to better prepare teachers to be accurate assessors of their students' reading ability.

National Survey of the Use of Test Data for Educational Decision-Making

In this chapter, Sheila Valencia presents the results of a nationwide survey of the uses of reading assessment data. This comprehensive survey drew from a sample of teachers and administrators which included schools of different student enrollment, location (rural, suburban, urban), and grade level (K-4, 4-8, 8-12). Valencia describes the content and scope of reading assessment in the United States, and provides a description of how statewide reading assessment fits into the broader reading assessment picture.

The survey reported by Valencia was conducted with three general goals: to obtain an accurate description of the scope and nature of testing in general, and reading in particular, in United States schools; to determine how reading tests and test data influence teachers' and administrators' actions; and to compare the actual and perceived impact of reading assessment.

Valencia uses the survey results to describe how much testing is being conducted, what types of tests are being administered, how test data are being used by teachers and administrators, how testing is influencing instruction, how teachers and administrators perceive the usefulness of test data for making instructional decisions, and how teachers and administrators perceive each others' uses of test data.

Statewide Reading Assessment: A Survey of the States

In this chapter, Peter Afflerbach presents the results of a nationwide survey of statewide reading assessment practice. Results indicate that 45 of 50 states use, or plan to use, statewide reading assessment. The survey describes how reading assessment currently is conducted at the statewide level. The summary includes information related to the type of assessment, the nature of the tasks included in the assessment, the grade levels at which students are assessed, the size of the student populations that are assessed, and the purpose of the assessment. Features unique to a particular state's assessment program are noted, as are recent innovations in statewide reading assessment.

Chapter 1

The Call for Assessment of Reading at the Statewide Level

The Call for Assessment of Reading at the Statewide Level

Peter Afflerbach, University of Maryland

This chapter examines several issues related to the statewide assessment of reading. First, factors contributing to the increasing use of statewide reading assessments are considered. Second, the variety of uses of the information gathered through statewide reading assessment is described. Next, issues related to the validity of statewide reading assessments are considered. Finally, the potential effects of assessment on reading curriculum and instruction are considered.

Factors contributing to the use of statewide assessment of reading

Helping students develop as readers has been a consistent goal of education in the United States (Resnick & Resnick, 1977; Applebee, Langer, & Mullis, 1988; Anderson, Hiebert, Scott, & Wilkinson, 1985). Statewide assessment programs in reading are considered by many to be useful gauges of progress towards this goal. As a result, statewide reading assessment has assumed several roles, including determining minimum competency in reading, identifying students' reading difficulties so that they may be remediated, and determining those reading program features which foster the development of reading ability (Fiske, 1988; National Commission on Excellence in Education, 1983; Wigdor & Garner, 1982).

A large percentage of statewide reading assessments consist of standardized, norm-referenced tests. Several factors contribute to the popularity (and the apparently uncritical acceptance) of standardized tests in the statewide assessment of reading (Airasian, 1987). Many American adults view standardized tests as fair, objective, and scientific. The tests are perceived as fair because all students take an identical test; and objective because test scores are fairly immune to bias introduced by teachers, principals, or parents. Tests are also perceived as scientific because they reduce the test-taker's performance to a numerical score. Statewide reading assessment is also considered efficient because it can provide reading ability data for each student at a particular grade level in a single testing session. However, the notion of "efficiency" of reading assessment may be more complex than is often acknowledged (Johnston, 1989).

A tradition of testing also contributes to the popularity of standardized reading tests. Standardized tests are linked symbolically to the maintenance of standards in education, and to traditional values. Thus, it is not surprising that recent calls for a return to basics in the curriculum and increasing school accountability are accompanied by statewide assessments, and their standardized testing components. In some instances, it may be argued that reading instruction program accountability for a school, district, or state has become reified as the administering of standardized reading tests.

In summary, the popularity of the current regimen of statewide reading assessments is attributable to several factors. Large scale reading assessment efforts are desired and trusted by a majority of American adults, as the efforts are perceived as ultimately leading to increased school achievement. It is the perception of many that the use of reading tests contributes to the maintenance (and perhaps the raising) of educational standards, specifically reading performance, and that tests do so in an objective, fair, efficient, and scientific manner.

Potential uses of statewide reading assessment data

The evaluative data gathered in statewide reading assessments is used for varied administrative, diagnostic, and selection and classification purposes. Administratively, statewide reading assessment results may be used to monitor the effectiveness of educational systems. For example, when reading assessment provides feedback related to the effectiveness of a particular reading instructional program, the results may indicate the need for changing or maintaining particular instructional programs. Statewide reading assessment information may also be used in decisions related to allocation of resources. Funding for instructional programs which appear to contribute to student achievement, and allocation of funds for schools and districts with relatively low achievement may be determined, in part, by reading assessment scores.

State education departments may use statewide reading assessment data to help establish a degree of control in the process of education at the local school district level. By holding school districts accountable for coverage of particular curricular content by assessing students' learning of that content, a state can achieve some standardization in the content of instruction. Additionally, mandating reading assessment may be part of a state's attempt to create and monitor minimum educational standards .

At both local and statewide levels, reading assessment results may be used to communicate school accomplishments to various publics (Airasian, 1987). Unfortunately, this may lead to the development of educational discourse in which test scores are considered synonymous with achievement (Koretz, 1989), or in which the vocabulary used to describe achievement is restricted to test scores.

Used diagnostically, statewide reading assessment results may influence instructional decisions, including the placement of students in reading groups or the selection of a particular reading instructional program. For example, a relatively low score on a statewide reading assessment may be used as one indicator of a student's need for remediation in reading. As part of a portfolio of student achievement, a test score may corroborate other indicators of students' reading ability (see Hiebert, this volume). Statewide reading assessment results are also used in prescribing instructional treatment and placement at the individual student level (see Valencia, this volume). In addition, statewide reading assessment results may be used in the determination of teacher accountability, in which student performance on statewide assessment is considered an indicator of teacher effectiveness.

The selection and classification of students according to statewide reading assessment results occurs in a manner which might be described as before, during, and after reading instruction. Before reading instruction, students' performance on statewide reading assessments may be used for placement in a particular reading program within a school or classroom. For those students performing above or below average, placement in gifted and talented or remedial reading instruction may be recommended. During reading instruction, statewide reading assessment results may indicate the ongoing effectiveness of reading instruction, and suggest areas of strength and weakness in both students' reading ability, and reading instructional programs. After reading instruction, students may be required to take "exit exams", which evaluate students' minimum competency in reading, or demonstrate that students are qualified for promotion or graduation.

Concerns with the nature of statewide assessments of reading

The potential for statewide reading assessment to impact on educational funding, classroom instruction, student placement, and subsequent learning is great. Thus, it is important to consider critical issues related to the tests which comprise many statewide reading assessments. Recently, the reading research community has raised several concerns related to standardized testing of reading. These concerns include the congruence between reading tests and current knowledge of the reading process, and the ability of the types of tests found on statewide reading assessments to accurately assess the interactive processes of reading. Additional concerns include the impact of reading tests on the reading curriculum, and the impact of testing on classroom teachers.

Given the frequent use and potentially widespread influence which statewide reading assessment results may have on decision-making related to reading instruction, the validity of the reading assessment instruments is an important concern. Critics of standardized tests of reading cite the lack of congruence between the current understanding of the interactive nature of reading and the nature of statewide reading assessment tasks.

Many of the instruments used in statewide reading assessment, and their ability to provide valid data related to students' reading ability, are suspect. While a comprehensive account of the potential weaknesses of standardized, norm-referenced tests which are found in most statewide assessments of reading is beyond the scope of this paper, several will be considered in the following section.

Reading is currently viewed as a dynamic process in which the reader interacts with written language in a particular social context to construct meaning (Anderson, Hiebert, Scott, & Wilkinson, 1985; Van Dijk & Kintsch, 1983). A purposeful reader uses appropriate prior knowledge and a coordinated set of processes and strategies, which often include questioning and inferencing strategies, to help in this construction of meaning. Readers often read texts which are complex and substantial in length. Depending on the reader's prior knowledge and purposes for reading, a text may be interpreted in several ways.

Do current methods of reading assessment reflect an understanding of the dynamic, interactive nature of reading?

Members of the reading research community do not think so (cf. Johnston, 1989; Valencia & Pearson, 1987; Wixson, Peters, Weber, & Roeber, 1987), and specific criticisms of reading tests are numerous. In contrast to our understanding of reading as a dynamic, interactive process, standardized, norm-referenced tests of reading are constructed to try to remove certain prior knowledge influences from the reading process. The texts found in standardized tests are often chosen so that their topics will be unfamiliar to most readers. Many reading tests assess the interactive process of reading as a set of discrete set of subskills, rather than a coordinated set of processes and strategies. In addition, only one "correct" interpretation of the texts included in reading assessments is allowed, as the purpose of reading the text is determined by the test constructor and test situation, rather than by the reader. The tests use short and contrived texts, the likes of which are found only in test booklets. The student-as-test-taker reads and answers questions in a social context unique to testing. Finally, statewide reading assessments fail to tap readers' use of strategies because the tests have a comprehension product (as opposed to process) orientation.

In summary, evolution in our understanding of reading as a dynamic, interactive process is in contrast to the lack of change in the way many standardized tests assess reading. Methods of reading assessment, if they are to be considered valid and useful indicators of reading ability, should reflect a refined understanding of the nature of reading. At the core of this problem are issues related to the validity of most statewide reading assessments. Regardless of the popularity of tests, lack of validity of statewide reading assessments may contribute to instructional practice and decision-making which are at best inappropriate for, and at worst harmful to, students.

The influence of testing on curriculum

A second concern related to statewide assessments of reading is the extent to which such assessments may determine the reading curriculum in schools. Consider the following teacher interview excerpt, which is taken from a recent study of teachers' methods of assessing literacy. In this excerpt, the teacher responded to a question which asked her to describe a typical instructional day in her 8th and 9th grade reading classes. The teacher is employed in a large, urban school district which places an extreme emphasis on standardized test scores:

- "...I have the eighth grade class and they're preparing to take the eighth grade state test...the Basic Skills Test...and the Iowa Test of Basic Skills...on the ninth grade they are prepared to take the TSP...the Test of Scholastic Progress...

- and they have to prepare for these skills...I have to teach them the skills that they need...follow the objectives...and prepare them to take the TBS (Test of Basic Skills) which is the exit exam...they're expected to be prepared in the class

- ...so we have to constantly have these objectives in mind as we teach on a day to day basis" (Johnston, Weiss & Afflerbach, 1989)

Given this description, it is difficult to imagine students actually reading. In the extreme, preparation for reading assessment may become the focus of instruction, as indicated by the above interview excerpt. Similarly, an emphasis on improving test scores may result in "teaching to the test", and avoidance or elimination of those reading skills or strategies which are not included on a particular reading assessment.

While the influence of high stakes assessments on the curriculum may not always be as pronounced as this (see Valencia, this volume), the nature of the assessment may influence curriculum in a more subtle (but as pervasive) manner. For example, a school district might select instructional materials (such as a basal reader series) whose assessment components most closely match those of the reading tests which are administered at the statewide level. If testing influences the curriculum, and the majority of standardized testing assesses reading as a set of distinct subskills, it should not come as a surprise that testing contributes to the continued skills approach to teaching reading that is followed in the majority of American elementary school classrooms.

The influence of testing on teaching

A third concern related to statewide reading assessment is the effect of assessment on the teaching profession. Use of statewide reading assessments and over-reliance on the evaluative data they provide may prove a hindrance to the increasing professionalism of teachers. Teachers must be "trusted" by administrators, parents, and the general public before their role in assessing students' reading abilities is more fully realized (Guba & Lincoln, 1983). Yet the continued emphasis on test scores works against the development of such trust, and contributes to the practice of overlooking, ignoring, or not seeking teacher-based evaluative information.

however valuable it might be (Hiebert, this volume).

Administering statewide assessments may also contribute to the development of an adversarial relationship between student and teacher, whether or not the teacher endorses the reading assessment which he or she must administer. Additionally, if the reading curriculum is driven by statewide reading assessment concerns, teachers may be forced to give up further control of their professional decision-making. For example, a basal reader system which is similar in content and format to a state's particular reading assessment may be used exclusively for reading instruction. Within this restricted instructional program, teachers may be required to adhere to an instructional timetable which guarantees completion of particular basal reader units, and prevents diversity of instruction.

Conclusions

The statewide assessment of reading is a popular practice. Such assessments are considered by many to be an accurate measure of progress towards educational goals. The statewide assessment of reading is representative of a tradition of large-scale testing in the United States, testing which is considered objective, fair, efficient, and scientific by many. The data gathered in statewide reading assessments is used for many purposes, including administrative, diagnostic, and selection and classification.

As popular as statewide reading assessments are, serious questions about the ability of the majority of the assessments to provide data related to the dynamic, interactive process of reading have been raised. In general, statewide reading assessment instruments have been slow in incorporating revisions which reflect an increased understanding of how reading "works".

There are several additional concerns related to many statewide reading assessments, especially those assessments which use a standardized, norm-referenced format. Statewide assessments of reading may influence the instructional materials and methods which are used in the class. Teaching to the test will lead to greater constraints on what is taught and how it is taught. Teachers' professional decisions about what to teach and how to teach it may be pre-empted by administrative or statewide curriculum directives which mandate instruction to prepare students to take tests, rather than to help students become better readers.

In summary, statewide assessment is a popular gauge of educational achievement in a subject which is traditionally valued: reading. However, issues related to the validity of statewide reading assessment, and the influence of assessment on reading curriculum and instruction suggest that the nature of reading assessment and the uses of assessment data should be carefully considered. Those who develop and use statewide reading assessments may be in a position to balance the need for measures of students' reading ability with the need for increased accuracy of the measures.

The result may be more effective statewide reading assessment programs, and the fostering of increased reading ability among students.

References

Airasian, P. (1987). State mandated testing and educational reform: Context and consequences. *American Journal of Education*, 393-412.

Anderson, R., Hiebert, E., Scott, J., & Wilkinson, I. (1985). *Becoming a nation of readers.* Washington, DC: National Institute of Education.

Applebee, A., Langer, J., & Mullis, I. (1988). *Who reads best?* Princeton, NJ: Educational Testing Service.

Fiske, E. (April 10, 1988). "America's test mania." *The New York Times Spring Education Supplement*, 16-20.

Guba, E., & Lincoln, Y. (1983). *Effective evaluation: Improving the usefulness of evaluation results through responsive and naturalistic approaches.* San Francisco: Jossey-Bass Publishers.

Jaeger, R. & Tittle, C. (1980). *Minimum competency achievement testing: Motives, models, measures and consequences.* Berkeley, CA: McCutchen.

Johnston, P. (1989). Constructive evaluation and the improvement of teaching and learning. Teachers College Record, 90, 509-528.

Johnston, P., Weiss, P., & Afflerbach, P. (1989). *Teachers' evaluation of teaching and learning in literacy and literature.* Technical Report to the U. S. Department of Education, Office of Educational Research and Improvement, Grant #G008720278.

Koretz, D. (1989). The new national assessment: What it can and cannot do. *NEA Today*, 7, 32-37.

Livingston, C., Castle, S., & Nations, J. (1989). Testing and curriculum reform: One school's experience. *Educational Leadership*, 46, 23-25.

Madaus, G. (1983). *The Courts, Validity and Minimum Competency Testing.* Boston: Kluwer-Nijhoff.

National Commission on Excellence in Education. (1983). *A nation at risk.* Washington, DC: U.S. Department of Education.

Resnick, D. P., & Resnick, L. B. (1977). The nature of literacy: a historical explanation. *Harvard Educational Review*, 47, 370-385.

Stiggins, R. (1988). Revitalizing classroom assessment: The highest instructional priority. *Phi Delta Kappan*, 69, 703-710.

Valencia, S., & Pearson, P. (1987). Reading assessment: Time for a change. *The Reading Teacher*, April, 40, 726-732.

Wigdor, A. & Garner, W. (1982). *Ability Testing: Uses, Consequences and Controversies*. Washington, DC: National Academy Press.

Wixson, K., Peters, C., Weber, E., & Roeber, E. (1987). New directions in statewide reading assessment. *The Reading Teacher*, April, 40, 749-754.

Chapter 2

Developing a Statewide Reading Assessment Program

Developing a Statewide Reading Assessment Program

Linda Hansche, Georgia State University

Reading assessment is not a new topic. It is, however, a topic under much scrutiny because many educators believe it has not kept up with current knowledge about the reading process. Within the investigation of reading assessment presented in this volume, this chapter focuses on how one state developed and continues to maintain a large-scale criterion-referenced reading assessment program.

Background

The foundation for the current assessment program in Georgia began in 1969. That year the State Board of Education appointed a blue ribbon panel called the Advisory Commission on Educational Goals. The panel included educators, parents, and members of the business community. Their charge was to investigate the educational system in Georgia and to develop goals for education in the 1980's. A year later, in 1970, the State Board of Education adopted their report. That report was the progenitor of the current student assessment program in Georgia.

In 1971, plans were formulated to developed a criterion-referenced assessment program to assess the new goals for education. By 1974 legislation was passed, called the Adequate Program for Education in Georgia (APEG), which required an evaluation program designed to systematically assess the educational goals. Since the criterion-referenced program was already under development, the legislation legitimized the efforts that were being made at the time. Thus, a criterion-referenced assessment program became the centerpiece for the state's educational evaluation plan. At that time, reading and mathematics were the major focus, with writing added recently.

In 1985 the Georgia Legislature passed a new educational reform bill called the Quality Basic Education Act (QBE). The QBE legislation reinforced the mandate for assessment in state educational programs. It called for the creation of an assessment program to measure a new state curriculum, the Quality Core Curriculum, which was also required by the law. The new program, called the State Item Bank, is being designed to assess multiple subjects across grades 1-12. The scope is to be much broader than the minimum competency focus of the present student assessment program.

Beginning the student assessment program

The student assessment program in Georgia is based on the concept of minimum competency and reflects only this type of assessment strategy. The current program began with a very specific purpose which was gradually expanded to encompass a broader need.

In 1976, the State Board of Education adopted the High School Graduation Requirements Policy 30-700 (HSGRP). This measure, which set new standards for high school graduation, required that students awarded a high school diploma in Georgia must have demonstrated competency in reading, writing, mathematics and problem solving.

To implement the new graduation policy, ten public school systems were selected to establish pilot programs. One of the charges to the systems was to identify those competencies which were necessary for adult life roles as learners, individuals, citizens, consumers, and producers. While the systems developed a curriculum, the Georgia Department of Education awarded a contract to Georgia State University to develop the assessment instruments. One of the instruments was to be designed to assess reading. Pursuant to that contract, the University established the Georgia Assessment Project (GAP) which acted then, and continues to act, as the test development agency for the state.

Beginning in August, 1979, the Georgia Assessment Project staff evaluated the information from the ten pilot systems and planned a procedure to a develop criterion-referenced test, called the Basic Skills Test, to be used specifically as one requirement for high school graduation. This development process included

- the development of content specifications for the reading competency area and a test blueprint;
- the development of a reading item pool to be used in constructing assessments for use in making pass/fail decisions about individual students;
- the development of a system for providing diagnostic information to individual students and their teachers about students' reading performance;
- the development of a strategy which would insure equated reading cut scores from form to form;
- the specification of a standard reading assessment score scale to be used in reporting student performance.

Soon after the HSGRP implementation was begun, an additional mandate by the State Board of Education for assessment of reading at several grade levels led GAP to expand its model for test development. The initial development process was modified to a generalizable criteria-referenced test development model for use at any grade level. A decision was made to maintain the original name, Basic Skills Test, commonly referred to as the BST, for the high school test. The additional reading assessments for other grade levels were termed the Georgia Criterion-Referenced Tests, or GCRTs. At the state level the whole program is referred to as the Student Assessment Program.

The model for the criterion-referenced test development and maintenance for the student assessment program developed by Georgia Assessment Project was designed so that Georgia educators provide input at every stage of development. As such, the process is not always time or money efficient, but more importantly it insures an process that directly reflects the Georgia curriculum and student needs. Within the model, the GAP staff acts as facilitators and technicians.

As facilitators, the Project staff recruits reading educators who are willing to become test item writers, reviewers, judges of bias, and item editors. As technicians, the Project staff provides the necessary training for educators to learn these various tasks, maintains the integrity of the process, and provides all the data analysis as well as any other statistical support for the program. The development model has worn well over the years and the basic model is still in use. The remainder of this chapter presents that model in some detail beginning with defining the content domain.

Content definition

Any good assessment tool requires clearly and carefully defined content. The process of content definition for the Basic Skills Test developed by the Georgia Assessment Project involved several procedures. The first step in the preparation of content descriptions for reading utilized the development work and recommendations of the ten pilot systems in which data were gathered to define the practical aspects of the High School Graduation Requirements Policy. Each of the ten systems had written, reviewed, and refined reading performance indicators based on the competency requirements specified in the HSGRP. In developing these indicators, a main emphasis was to match them to reading curricula. Judgments were also solicited about the importance of standards for a minimum-competency requirement for high school graduation. This judgmental process typically involved teachers, curriculum specialists, parents, and other representatives from the local business community who were asked to rate the importance of the performance indicators, revise them, and suggest additions and/or deletions.

When the data were organized, a validation workshop was conducted by GAP with participation at this stage limited to educators. The workshop participants were asked to

- sort the indicators into groups or categories;

- identify a parsimonious set of categories and reclassify indicators if necessary;

- make recommendations for eliminating redundancy among the indicators;

- review and rate each indicator according to its importance in the reading curriculum and for high school graduation; and

- recommend appropriate strategies for assessing the performance indicators.

When the classification, review, and rating tasks were completed, participants were asked to discuss their concerns for various assessment strategies. A major concern was to provide an assessment strategy which focused on the application of skills in situations like those students might reasonably be expected to encounter when using their reading skills in classrooms, everyday life, and work settings. Additional concerns focused on the importance of avoiding a simple measure of reading vocabulary or factual knowledge and the need for a reading assessment instrument that could be administered consistently and fairly across the state to insure validity and reliability.

After the workshop was completed, the next step was to formalize the reading performance indicators into assessment objectives. The systematic procedure was based on a conceptualization that each objective must specify three dimensions. These dimensions included (1) the response required of the student (e.g., to identify, to interpret); (2) the specific reading content (e.g., sequence of events, relevance of data); and (3) the social context of the item. The last dimension, social context, was included to avoid situations assessing facts or vocabulary and to help direct the focus of reading assessment toward life roles. More specifically, three social contexts were defined. Items written for the academic context utilized printed materials typically encountered in classroom or other instructional situations. Items written to reflect an everyday context were based on materials related to personal interactions, label information, set of directions and the like. Items in an employment context involved materials related to work situations such as application forms, employee insurance policies, and manuals.

Approximately two hundred performance indicators were reduced to twelve broad reading objectives. With the help of reading specialists, the GAP staff began preparation of a technical document to define each objective by creating item specifications. Specifications for each objective include information essential for describing the content. Terms are defined and explained. Ranges or limits of content are outlined. Item specifications include suggestions for appropriate stimulus material, e.g., book reviews, narratives, labels, and types of acceptable graphics. More recent versions of the item specifications also include a section on strategies item writers might use for creating effective distractors, or incorrect options. A final section for each objective presents several sample items.

When the item specifications were completed, the final step in developing the reading assessment objectives focused on validation of (1) the appropriateness of the assessment objectives relative to the original set of competencies, an aspect of construct validity; (2) the match of the objectives to curriculum guides, an aspect of curriculum validity; and (3) the perceived importance and emphasis placed on skills and concepts described by the set of objectives, an aspect of content validity. Three procedures were used to provide this evidence. The first procedure was the review and evaluation of objectives by reading professionals. The second procedure was a survey based on a sampling of the state's teachers, school board members, and parents. The third procedure was an opinion survey of tenth-grade high school students who took the field-trial test, which sought commentary on difficulty of the material and the relevance of such a test.

Results from each of the three procedures indicated strong agreement that the objectives were clearly stated and defined and that the objectives assessed the required reading competencies. Those surveyed also strongly agreed that the content of the objectives should be assessed, indicating that the initial set of learner competencies for reading was still useful. Upon acceptance of the set of reading assessment objectives, item development became the next focus.

Item development

An essential aspect of Georgia's reading assessment program is its heavy reliance on Georgia educators. Most educators who are involved with the development process at GAP are first trained to be item writers. This experience gives consultants the foundation and information they need for participation in subsequent assessment activities. In many ways, the item writer training program is the backbone of the Georgia Assessment Project model. Recommendations for consultants to work with GAP are solicited from the Georgia Department of Education, curriculum directors in each school system, and other GAP consultants. While participants are most often classroom teachers, curriculum specialists and administrators with expertise in the field of reading are also invited to collaborate.

GAP item writer training workshops are highly interactive. Formal presentations by GAP staff alternate with participative activities and questions or other types of input from the consultants are encouraged. Time is divided between small and large groups and include discussion and critiques of items. The method of training is labor intensive, but it produces items of appropriate quality and focus which are based on item writers' experiences with Georgia students and curriculum.

GAP provides a basic item writer training workshop, and in addition offers an advanced item writing workshop at which previously trained writers reinforce and refine their skills.

Basic Item Writer Training. In the first component of the training session, assessment program requirements and item terminology are described and defined. Consultants are given an in-depth coverage of the reading content to be assessed. Objective content, or item specifications, appears in a special document called an item writing guide. The overall structure of an objective is explained, including the content, response, and context components. For each objective, assessment characteristics are presented and discussed. Reading terms are defined to ensure that the item writers, whether or not they agree, understand the meaning and the intent of each objective. Participants are advised about any restrictions on types of items that may be written for an objective as well as any content restrictions that may be included. Sample items appearing in the writing guides are explained in detail, and additional sample items are presented to further illustrate how an objective or parts of an objective may be assessed.

When the item writing guides have been explained and fully discussed, the training shifts to the characteristics of good multiple-choice assessment items including a special session on bias. Participants are shown examples of both well-written and flawed multiple-choice items. They are asked to react to each item. After the group has examined numerous examples, a GAP staff member presents a list of guidelines for writing good multiple-choice items which is contained in their training materials. Guidelines include statements like "No one option should be a subset of any other option"; "If there is a passage, it should be necessary to answer the item(s)." Each guideline is illustrated using acceptable and unacceptable examples. After all the guidelines have been considered, participants are presented with another set of sample items. This time, items exemplify specific flaws which have been discussed. Participants are asked to point out any problems and then to provide alternatives for correcting them.

At this point in the item writer training sequence, it is expected that the participants possess the basic knowledge needed to begin writing assessment items. The next phase requires participants to work independently to draft several items. Participants are instructed to concentrate on those types of items that do not accompany reading passages, since passage writing is addressed later in the training sequence. A GAP staff member leads a critique of each item (anonymously submitted) reinforcing the material presented earlier in the workshop.

Trainees are next introduced to passage writing. To stimulate ideas for passages, the group brainstorms possible topics. For each topic idea that is generated, participants are asked to think of related topics and of possible approaches to the suggested topic. The need to avoid certain inappropriate topics or controversial issues is also a part of this activity.

Following the brainstorming activity, participants are given time to choose a topic and to draft a passage of their own. Passages are required to be original, realistic, and accurate. GAP staff members are available during this time to provide individual feedback. Resources are available at the workshop, including encyclopedias, magazines, and text books on various topics. The passages drafted by the participants are transferred to overheads and again critiqued by the group. The discussion is focused on characteristics of good writing in general, and specifically on the elements of prose that make a text cohesive. Trainees are reminded that their passages should not be merely mechanically constructed devices for assessing a student's reading; they should provide the examinees with accurate, well-written, interesting text.

Next, participants write items to accompany their passages. The ensuing critique focuses on both the passage and the items. Passage dependence and item independence are compared. Passage dependence means that the reader must use the text to respond correctly to an item. Item independence means that correctly or incorrectly responding to one item does not influence chances of correctly or incorrectly responding to another item. The GAP staff also provides advice on how to modify or revise passages in ways that will provide for items which are appropriate according to the item specifications, that will allow for creation of better or more varied distractors, or that will allow additional items to be written for that same passage.

Advanced Item Writer Training. The advanced writer training workshop differs from the basic workshop in several ways. Although staff members are available at all times for individual feedback, participants are encouraged to assist each other and discuss passages and items under development among themselves.

The advanced training workshop includes an intensive session on passage construction. Sample passages are examined and systematically analyzed with regard to the flow of ideas and the relationships among them. The purpose of this session is to aid writers in creating passages that are well-structured and which allow for a maximum number of associated items.

A second major component deals exclusively with distractor strategies. Although the importance of distractor strategies is addressed in the basic item writer training workshop, it is only at the advanced workshop that the topic is treated at length. In addition to the section in the item specifications, writers are presented with material specifically created for that particular workshop listing characteristics of good distractors by objective and strategies for creating them. The participants are shown examples of actual field-tested items. They are asked to speculate on which distractors were most attractive or least attractive. Then they are shown the actual percentage of examinees choosing each distractor. This activity helps make the writers aware of the factors that may contribute to the attractiveness of a distractor. Each trained writer, whether basic or advanced, is given an item writing assignment that specifies the number of items required per objective based on the specific item bank needs. Each item writer is also asked to make recommendations as to the relative weight each objective should have on an actual test form.

This information is returned to GAP along with the completed items and used later for weighing the selection of items used in constructing assessment instruments. After items are written the item review process begins.

Item review

The next major component of the GAP assessment model is the process of item review and editing. The majority of reviewers are selected from the pool of trained item writers, representing a range of grade levels and job titles. If, for example, third grade items are being reviewed, some second and some fourth grade teachers are typically included in addition to third grade teachers and early childhood curriculum specialists.

The first part of each item review workshop involves a re-examination of the characteristics of good multiple-choice items. Reviewers are provided with information about reviewing items for potential sources of bias. Written guidelines describing possible types of racial, cultural, gender, and task or situation bias that might occur in an item or in a set of items are discussed at length.

When the reviewers are acquainted with the item review procedures, they are assigned to small groups consisting of two to four participants including a Project staff member or experienced item writer/reviewer. The group begins the job of editing each test item. Items are categorized as "good", "omit", or "hold" (the latter category usually needing some sort of verification of information). During the review process, any instances of bias identified in an item or a set of items are recorded on a special form. Whenever possible, an offending item is "fixed" after the bias notation is made. For example, a reference to an angry person who has red hair would be modified so that the hair color is not stated. The bias notations provide feedback to the staff for subsequent item writer training sessions.

When a review group has examined an entire set of items one by one, they are required to go back and consider the set as a whole. Reviewers are asked to tally names as a check for ethnicity and male/female distribution. While there is no set quota, the decision to change a name or role is based on reviewer judgement of a balanced representation of various groups of people. At this stage, reviewers also check for coverage of content as well as for any biasing elements in the set, such as the over-representation of urban experiences, or the portrayal of females mostly in stereotypic roles.

Following the item review workshop, "hold" items are reviewed again in-house and are often salvaged by verifying passage and/or item information for accuracy. Items labeled "good" and the verified "hold" items are prepared for field-testing. The items are entered into a computer system used to maintain item banks. Each item undergoes an additional technical review by Project staff as well an outside consultant before a final version is prepared for field-testing.

Field-testing

All items are subjected to a field-test procedure before they become part of an item bank. Items to be field-tested are administered in intact forms whenever possible. This means that items are field-tested in the same book at the same time the operational form is being administered. The field-test sections are not identified as such, and care is taken to ensure that the field-test items do not differ significantly in format from the items in the operational sections. There are, however, instances where experimental items and/or formats are tried out and consistency cannot be achieved. While operational sections of the test are identical for all students, each of the field-test sections contain a different set of trial items. In grades where a new reading assessment is not routinely produced, field-test items appear in a supplemental booklet accompanying the operational form and are administered during the same testing period.

To prepare field test forms, status of the current item bank is reviewed. The GAP test development specialists search the pool of items recently written. Items to be field-tested are selected based on bank needs. The items are then parceled out into different test forms, with care being taken to achieve a balance of topics, item types, objective content, gender, and key balance whenever possible for each form. Before test forms are printed, a content consultant, usually a classroom teacher who is an experienced item writer and reviewer, verifies items one last time to ensure that each item does indeed match the objective it was intended to assess, that each item has only one correct answer, and that any other flaws are discovered and corrected.

Stratified random sampling is used to distribute the various field-test forms to each school system and/or classroom. First grade students are given scorable answer books; all other students use a separate answer document. Item data are analyzed by GAP personnel using a specialized program which provides both traditional and Rasch statistics. These are used to determine which of the field-test items are acceptable for use on an operational form. In addition, all items must pass a bias review.

Bias review

One of the most important features of Georgia's assessment program is its commitment to producing tests that are as free as possible of bias. In addition to emphasizing bias issues at both item writer training and review workshops, GAP conducts a special workshop at which items written for pass-fail assessments are examined specifically for bias. A bias review workshop is scheduled after field-test data are available. A special standing bias review committee meets to examine the items for bias. Committee members represent all levels of administration and instruction as well as various cultural groups and regions from around the state. Members serve a three-year term.

At the beginning of the workshop, reviewers are provided with images of all items exactly as they were administered along with statistical information on performance of black and white samples of students. Until recently, performance of male and female students was also provided as a source for potential bias as well as a comparison of regional data. Those analyses consistently have shown no significant differences between these groups within the state and data are no longer routinely provided.

Black-white data for bias include (1) item p-values (percent selecting each option), (2) adjusted item difficulty based on the Rasch analysis, and (3) a plot showing the relationship of item difficulty to student ability. Accompanying the data and the item images is material describing potential biasing elements in four categories: slurs, stereotypes, task requirements, and erroneous group representations.

After an explanation and review of the task and the materials, reviewers examine items and the accompanying data. Each item is reviewed individually by at least four committee members. While reviewers are asked to record any instances of bias, they are asked to pay special attention to those items that have been identified as statistical outliers, i.e., those items that appear to be far more difficult than expected for one group of students than for another. It is this differential item functioning (DIF) that is an indicator of potential bias. Reviewers are asked to examine outliers carefully and judge whether or not the difference is a reflection of a biasing element in the item content or presentation. They are also asked to note any technical flaws they find, even if these do not necessarily reflect bias. After reviewing the items individually, the reviewers are asked to share their concerns in an open discussion. They are asked to make recommendations that may become part of future item writer training sessions.

Following the bias review workshop, the comments of individual participants are compiled. Any problem items are, if possible, revised and field-tested again. Items judged irreparable for any reason are purged from the item bank. The remaining items are used to construct an operational instrument.

Construction of operational assessment forms

The process of selecting those items from the bank that will appear on an operational assessment form comprises three steps:

- development of a content matrix,
- preliminary selection of items by GAP staff, and
- final selection of items at an item selection workshop.

Before items for a first operational form can be selected, the number of items representing each objective must be determined. In making these decisions, GAP carefully considers the content weighing recommendations collected from consultants at item writer training workshops and item review workshops. Using these guidelines, GAP test developers create a content matrix which defines a target number of items for each objective based on an average of the recommended weights.

When the content distribution is finalized, preliminary item selection is conducted in-house. The preliminary selection of items by GAP staff is necessary because of the number of factors that must be considered. First, a set of items is selected from two previous forms to serve as an equating link. Approximately 30-40% of the items on any one operational form are link items. Of these, half are links to the most recent operational form and half are links to the operational form that preceded the most recent one, provided the form is not a first or second edition.

Stringent link item requirements are necessary to insure test reliability from form to form. The link items are used to equate the new form to previous forms and to ensure that the bank of items remains stable over time. In other words, the link items provide the basis for equating the difficulty of form A with that of form B with that of form C so that a comparison of student performance across forms is possible.

Once the link or overlap items have been chosen, others are selected to complete a preliminary set of items reflecting the targeted content balance. Although the second set of items must also meet certain statistical requirements, those requirements are less stringent than for the link items.

For the entire set of items, including both overlap and new items, the average item difficulty within an objective should be approximately equal to the average difficulty for that objective on previous forms. The reading passages included should reflect a variety of topics and an adequate balance of keys, male and female representation, and ethnic representation.

After a preliminary set of items is selected, experienced consultants attend an item selection workshop. Each participant in the workshop is provided with the set of preliminary items and field-test statistics for each item. Other materials provided for evaluating the set of items include a copy of the item writing guides, a checklist of considerations for item selection, a handout describing potential sources of bias, and a content matrix. The content matrix shows, for each objective, the number of items on previous test forms and the average objective difficulty, the number of items recommended, and the target number of items that should appear on the new form. In addition to the materials provided to individual participants, the entire item bank is on hand so that the participants may choose replacement items for those they eliminate from the preliminary selection.

Participants are asked to study the preliminary set of items and their accompanying statistics. They are asked to evaluate each item with regard to the following questions:

- Does the item have one and only one correct answer?

- Does the item assess the objective it was intended to assess?
- Is the item free of technical flaws?
- If the item accompanies a passage, is the passage required to answer the question?
- Is the topic of the passage current, accurate, appropriate, and interesting?
- Is the reading level of the passage appropriate for the grade level?
- Do item statistics appear to be within range?

After the participants have reviewed the items individually, concerns about items are discussed at length. If participants agree that a passage or an item is unsatisfactory, that item is eliminated from the set, and the group selects a replacement passage or item from the bank. When all desired replacements have been made, the group evaluates the new set of items as a whole. The set of items should reflect the required content distribution, a balance of male and female roles, and a variety of names, topics and situations; an acceptable range of difficulty; and a balance of answer keys. If the set is found to be imbalanced on any one of these parameters, further substitutions are made until the desired balance is achieved. The final selection must reflect all the necessary requirements while still utilizing statistically good-fitting items. Once the set of items is finalized, diagnostic information is generated.

Diagnostic workshop

One benefit of a criterion-referenced assessment is the potential for providing student feedback. Individual reading score reports for the GCRT include specific statements about a student's area(s) of strengths and weaknesses within the limits of the content. The content and wording of each diagnostic statement, as well as the level of student performance that warrants them, are determined by classroom teachers at special diagnostic workshops. A diagnostic workshop is held after each new operational test form has been prepared, but before it is administered. Workshop participants are provided with images of all the items that appear on the new form. Accompanying each item are field test data, including the Rasch item difficulty which has been adjusted to the bank, and p-values. A scattergram is also provided showing the relative positions of the items on a difficulty scale of all items for that objective.

Considering the content characteristics of the items by objective, participants decide whether the set of items should be subdivided for the purposes of diagnostic statements. Consider, for example, a set of items that requires the examinee to identify the main idea of a passage. For some passages, the main idea is found in the first sentence of the passage; for others, it is found elsewhere. If there are several "first-sentence" main idea items, the participants may choose to consider that item type separately for diagnostic purposes. By selecting and grouping "first-sentence" main idea items, a specific diagnostic statement may be created to print for those students who do not respond correctly to the items in that subset.

After determining whether one or more subgroups of items are useful within an objective, the participants determine how many items from each subgroup (or objective as a whole) an examinee should be allowed to answer incorrectly without receiving the diagnostic statement. GAP makes a strong recommendation that perfection, i.e., four out of four items correct, is usually not desirable. The intent of the assessment is not to produce diagnostic statements for any and all students who might need help, but rather to be careful that students who receive the statements in fact probably do need further instruction. There are many other important aspects of reading proficiency needing instruction beyond the skills assessed; the recommendation toward conservatism is made in an effort to minimize unnecessary skill level instruction.

As the content and wording of each diagnostic statement is determined by the participants, they are reminded that each statement should be meaningful not only to teachers, but to students and their parents as well; thus, it is important to avoid technical words or phrases. A diagnostic statement, for example, might read "You may need additional instruction in determining the meaning of unfamiliar words using context clues." After the workshop, the results, including statements and rules for generating them, are sent to the scoring vendor, where a computer program is developed to produce the appropriate diagnostic statements on student score reports when they are printed following administration and scoring.

Equating

GAP uses the Rasch model, a single parameter item response theory model. This model generates estimates of both the difficulty of an item and the ability of the students who attempt the item. An advantage of the Rasch model is that it provides sample-free difficulty estimates for items and item-free ability estimates for examinees.

The purpose of equating tests is to make scores from different test forms comparable. The observed difficulty of items is determined by the abilities of the students who attempt them. Typically from one administration to another, neither the pool of students nor the set of items is the same. Therefore performance cannot be compared unless the various sets of items, i.e. test forms, are placed on a common scale. Throughout the test construction process an attempt is made to insure that test forms match in content and item characteristics. However, the process is not perfect. Through the technical process of item linkage and equating, any inequities in forms can be corrected. The raw scores are adjusted for any difference in the item difficulties so that students are neither penalized nor rewarded as a result of the particular form of the test they took.

In the beginning stages of development of a new item bank, a common origin for the bank difficulty scale must be defined. This common origin is generally defined as the mean difficulty of a selected set of field-tested items, usually a single form from the first set of field-test forms. All subsequent field-tests are adjusted, or equated to this common origin.

After the bank is initially scaled, each new operational form must be equated to the bank. To equate test forms, the process involves the use of a group of overlap, or link, items as discussed in the section on construction of operational forms. The performance of students on a common set of link items reveals any differences in the relative difficulty from previous forms to new form. To compensate for any inequities, a constant is computed to adjust the difficulty values of the new operational form items to the bank scale. By adjusting the difficulties and equating forms to the bank scale, performance can be compared across administrations. A further step in comparing performance is reflected in the standard setting procedure.

Standard setting

Like other criterion-referenced assessments, the Georgia program provides information that can be interpreted with regard to a specific standard of performance be it via diagnostic statements or pass-fail status. Criterion-referenced assessments by design are not intended to compare examinees with regard to a range of proficiency, but only to discriminate between those examinees who have reached a required level or standard of performance and those who have not. Rather than setting a high standard for achievement, the minimum competency concept used in Georgia focuses on basic or essential skills and the standard setting procedure reflects this orientation. State standards are set for those assessments used to determine a pass/fail status.

The Georgia Department of Education is responsible for conducting each standard setting workshop. Participants in the workshop are selected by the State Department of Education and typically include teachers, principals, school system superintendents, curriculum directors, and parents and members of the business community in the case of the Basic Skills Test. Various geographical areas of the state are also represented.

The most difficult issue related to the concept of a minimum competency assessment is that of arriving at a specific standard of performance that represents "minimum competency." This means setting a cut score that will determine whether a student passes or fails the test. Ideally, students who have minimal knowledge of reading will pass, while those who do not possess the minimal skills will not pass. Since the results of such a decision will necessarily have profound impact on individual students, their parents, and schools, much effort was spent investigating the relative merits of different standard-setting procedures. The procedure eventually chosen involves collecting data from judges who evaluate each test item with regard to how a minimally competent student would be expected to perform on the item.

At the workshop, participants examine each item and estimate the probability that the item will be answered correctly by a minimally competent examinee. The phrase "minimally competent" is defined for the participants at the beginning of each workshop in a way that is relevant to the purpose of the specific test. For example, since passing the high school Basic Skills Test is one of the requirements for high school graduation, at the BST standard-setting workshop a minimally competent examinee is defined as a student who exhibits basic reading proficiency at a level sufficient to warrant a high school diploma. To further assist participants in understanding their task, they are told that they may arrive at the probability values for an item by thinking of a hypothetical group of 100 students who are minimally competent in reading and then determining how many of those students they would expect to answer the item correctly.

When the participants have individually judged the probability values for each item, the p-values for the group are averaged and given back to them for re-evaluation. At this point, the participants are provided with actual student performance data from a previous administration of each item. This step is especially important since the first judgments were made strictly on the basis of intrinsic content characteristics of the items and participant judgement.

After studying actual student performance data, the participants re-evaluate the mean values assigned to each item. If desired, they may revise their initial probability estimates. The revised estimates are once again averaged and presented to the participants.

At this point in the workshop, the emphasis shifts from evaluating specific items to evaluating total test scores. From the probability estimates obtained earlier in the process, a tentative cut score is determined. Participants are then shown an actual distribution of test scores based on past administrations. This allows an estimate of the number and percentage of students who would be predicted to fail the test if the tentative cutoff score were adopted. Again, the participants are given the opportunity to evaluate their initial judgments and revise them if they choose. A final recommendation is then made.

The recommended cut score is submitted to the State Board of Education. It is the State Board that sets the actual required score for each test. Those scores or standards are periodically re-evaluated in light of changes in curriculum and instruction or in the student population. When a new standard is indicated, a new standard setting workshop is conducted and the recommendation is again submitted to the State Board for their approval. With the setting of a standard, the test development cycle is completed.

Future direction

The assessment model developed by Georgia Assessment Project for the Georgia Department of Education has proven itself effective for the past 10 years. It takes approximately two years and one hundred consultants to develop a new first-time operational assessment tool. Depending on the grade, new operational forms are developed anywhere from three times a year to once every three to five years. The constant contact with and input from state educators keeps the assessment development as current as possible, given the original charge and set of objectives.

Generally, the many people involved in the development of the CRTs are proud of what has been accomplished. However, it is common knowledge that within the "givens", (i.e., a large scale, one-right-answer, paper-and-pencil multiple-choice assessment), much of what we know about students' reading ability is not, indeed cannot, be assessed. The CRTs focus on skills and products, not thinking and process.

In Georgia, new ways of looking at reading assessment as a means of improving reading instruction are being explored. Most educators are aware that more testing under the guise of accountability will not improve education. As of this writing, no official action has been taken by the state. However, an independent group of reading professionals committed to promoting more effective reading instruction is examining the current assessment program. The only given they are working with is that the state must have and will maintain a large-scale reading assessment program.

As noted elsewhere in this volume (Teale, Hiebert, and Valencia), the mandate for change in the manner in which reading is assessed is apparent and can no longer be ignored. A most important issue is what type of assessment can be used to maintain the integrity of the reading process. A related issue is how to maintain that integrity while accurately measuring the reading process. Yet another issue is how to produce an instrument that meets the demands of a large-scale assessment program. The initiative for change is yet at the grass roots level in Georgia; we at Georgia Assessment Project plan to be a part of making that initiative a reality and hope to be ready to meet the challenge when it is issued.

Chapter 3

Issues in
Early Childhood Assessment

Issues in Early Childhood Assessment

William H. Teale
University of Texas, San Antonio

Statewide assessment of young children's reading is certainly alive in the United States. Afflerbach's recent survey showed that 16 states currently assess the reading development of children in first or second grade (see his chapter in this volume). In addition, four states screen kindergarten children for reading readiness as part of an assessment program. In most instances states use formal, standardized measures (either criterion-referenced instruments that are part of a minimal competency testing program or norm-referenced reading achievement tests) to accomplish this assessment. Therefore, for all intents and purposes, statewide assessment of young children's reading, therefore, means statewide testing.

The practice of statewide testing of young children's reading has been closely examined in many parts of the United States. For example, in 1987, North Carolina passed legislation that replaced standardized testing of reading in first and second grades with "developmentally appropriate individualized assessment instruments." North Carolina's Department of Public Instruction responded by implementing measures that sample student performance in oral language, orientation to print, listening and silent reading comprehension, reading strategies, writing, and integrated communication skills (Division of Communication Skills, North Carolina Department of Public Instruction, 1989).

Concerned about the stress of mandatory testing on first graders, the Arizona legislature passed a bill in 1988 that limits testing to a sample of number of students. Beginning with the 1988/89 school year, Mississippi eliminated standardized testing of kindergartners because teachers were using the tests as curriculum guides. And as recently as the summer of 1989, the Texas legislature eliminated minimal competency testing in reading with students below the third grade level.

Professional organizations and policy groups have also addressed the issue of standardized testing of young children. The National Association for the Education of Young Children (NAEYC) and the National Association of Early Childhood Specialists in State Departments of Education (NAECSSDE) have warned that pencil and paper tests incorrectly brand some four-, five-, and six-year-olds as failures. They also warn that highly formal testing procedures are inappropriate for many young children (NAEYC, 1988; NAECSSDE, 1987).

Right From the Start, the report of the National Association of State Boards of Education's (NASBE's) Task Force on Early Childhood Education, agrees with these concerns and recommends widespread review of standardized testing programs as well as the development of new approaches to documenting and reporting young children's learning and achievement in areas like reading (NASBE, 1989).

In *Literacy Development and Prefirst Grade*, the International Reading Association, the National Council of Teachers of English, the Association for Childhood Education International, the Association for Supervision and Curriculum Development, the National Association of Elementary School Principals, and NAEYC, expresses concern that the pressure to achieve high scores on tests has led to undesirable changes in the content of kindergarten programs. This statement recommends the using developmentally and culturally appropriate evaluation procedures. It also recommends informing the public about the limitations of standardized measures of prefirst graders' reading.

The focus on standardized testing of young children exhibited by various states, professional organizations, and policy groups shows that the United States is rethinking early childhood reading assessment. It is a particularly opportune time to do so. Research on early literacy learning during the past decade has advanced the field considerably (Sulzby & Teale, in press). This research has led to important advances in curriculum and instruction (Strickland & Morrow, 1989).

With such developments comes the need for assessment that supports and advances the goals of reading programs in early childhood classrooms. This chapter examines early childhood reading assessment and proposes alternatives to its current state. Its main point is that teaching and assessment can be brought together in quality literacy programs for young children. It also discusses challenges that must be met if we are to succeed in using developmentally appropriate assessment programs in our early childhood classrooms.

Currently, standardized measures that statewide testing programs promote certain instructional activities that can detract from the acquisition of reading skills. Important aspects of early literacy development that the current literature identifies in the often go unassessed and thus are often absent from school curricula or underemphasized in actual classroom practice. Instead of standardized tests, I propose using informal and observational techniques to obtain information about children and to link instruction and assessment in the classroom in a positive way.

Concerns about Current Statewide Reading Tests for Young Children and the Influence of Statewide Tests on Instruction

A growing body of research indicates that testing shapes in various curriculum areas and that the overall effect of this state of affairs is negative rather than positive. Tested areas of subjects are emphasized in instruction at the expense of untested areas (Darling-Hammond & Wise, 1985). Some elementary teachers even take instructional time away from subjects that are not tested in favor of ones that are (Salmon-Cox, 1982, 1984). As Valencia reports in this volume, teachers use information from statewide reading assessments to diagnose the needs of individual students and to set instructional goals; purposes for which the tests were never designed. Because overriding emphasis is placed on 'basic skills,' higher order thinking, reading, and writing goals frequently receive little attention in the curriculum when minimum competency tests are the measures of achievement (Shepard, 1989). It has also been found that test-oriented instruction drives many good teachers out of the profession and "deskills" a number of others (McNeil, 1988).

These types of findings with older elementary students also appear in early childhood classroom when schools use testing approaches to reading or reading readiness assessment. The overall nature of the reading curriculum, as well as the day-to-day instructional interactions and activities can be affected.

An example from Texas illustrates the point illustrated above. The Texas Educational Assessment of Minimum Skills (TEAMS test) is a multiple-choice minimal competency test of reading, writing (the writing test includes a writing sample at grades 3 and above), and mathematics that was, until the Texas legislature eliminated the first grade test in 1989, given to children in alternating grades beginning with grade one. Reading skills assessed in the early grades included:

- main idea
- sight word recognition
- compound words (first grade only)
- context clues
- word structure
- phonics
- specific details
- sequencing events
- predicting outcomes
- and table of contents (third grade only)

Exactly how does this early grade reading assessment influence the reading curriculum? At the beginning of each academic year in one south Texas school district, teachers are given a book published by a less-than-major publisher located in a rural Texas town. The book consists entirely of pages of items like those found on all the TEAMS subtests. In essence, it is a testing workbook. These exercises are not put into larger reading or writing contexts; the pages are merely for practice on items as similar as possible to the ones found on the TEAMS test. Teachers are encouraged by the school district to use these books for about 15 minutes per day, up until the test is given in February or March. This means that these schools spend approximately 30 hours of class time, five complete days of the school year (the equivalent of four weeks of instructional time typically devoted to the language arts in first grade), on this work rather than on actual teaching of reading and writing.

In addition, three to six weeks before the test is given (depending upon the school district), preparations for the test intensify. Instead of 15 minutes per day, the reading and language arts supervisors and most of the building principals in one district suggest that teachers spend "at least 30 minutes each day preparing students" for the upcoming test. In February, 1989, numerous schools in the San Antonio are also held Test-Buster Rallies or TEAMS-Buster Rallies. In other words, enormous amounts of time were devoted to practicing for the test or 'psyching students up' to take the test. Clearly, instruction is directly affected by the tests. In fact, in many Texas schools it can fairly be said that the TEAMS test has become a blueprint for reading instruction.

Why do statewide reading tests have such a marked effect on what teachers have students do in the classroom? Because administrators and teachers perceive that these tests are used to make decisions about their success as professionals, these are high stakes tests (Madaus, 1988). So long as they are perceived in this way, their content and format will be translated directly into the classroom practice.

Such a testing-teaching relationship in reading can be particularly deleterious to young children for two reasons. First, because assessment is accomplished almost exclusively through formal testing procedures, the types of classroom activities engendered by assessment programs may be inappropriate to the developmental characteristics of young children. Second, the content of a statewide reading assessment program can overlook significant aspects of young children's literacy knowledge and behaviors. As a result, the content of the curriculum can suffer.

Developmentally Inappropriate Influences of Tests on Early Reading Instruction

Children from four- to seven-years old need a reading curriculum fundamentally geared toward promoting *knowledge* (cultural, social, and literary knowledge) and developing reading *strategies*. They need experiences with *purposeful* reading and with writing to a *variety of audiences*. They need to *discuss* and otherwise *respond* to literature in ways that promote higher level thinking. In short, they need to be involved in a "hands-on" approach to literacy in which reading is a problem-solving activity in classroom and life experience (Strickland & Morrow, 1989).

Curricula guided by standardized tests, however, lead children in other directions. Reading tests are not designed to hold children's interest. They tend to contain reading passages different in length and in kind from those most useful for instruction in the early childhood classroom. Standardized measures used in statewide kindergarten, first, and second grade assessment programs focus almost exclusively on component skills like those tested for on the TEAMS test. They rarely integrate these aspects into the skilled act of reading. As a result, young children can be expected to spend inordinate amounts of classroom time completing workbook pages, ditto sheets, or computer programs that mirror testing tasks. Young children are not well-suited for extended periods spent filling in worksheets.

Children this age learn best through active involvement in tasks and through social interaction with the teacher and peers. So, by promoting activities in which isolated children engage in pencil and paper exercises on isolated skills, standardized tests of children's early literacy development actually play to the developmental weaknesses of young children instead of capitalizing on their learning strengths. This is especially true for children who are considered at risk for failure in reading. Young children need reading instruction patterned more on an apprentice model of learning, a metaphor that has appropriately been applied to the overall experience of language development (Miller, 1977) and that serves well as a way of envisioning a productive approach to classroom teaching.

Content Problems of Reading Tests for Young Children

The past 10 to 15 years of research on young children's early reading has legitimized the concept of emergent literacy as a way to conceptualize the period of development from birth to the time when children are able to read conventionally and fluently (Strickland & Morrow, 1989; Sulzby & Teale, in press; Teale & Sulzby, 1986). Emergent literacy recognizes that prior to conventional reading and writing, children develop knowledge about literacy and engage in literate behaviors. These conceptualizations and behaviors are extremely important aspects of literacy ability and continued learning, and they develop in predictable ways toward conventional literacy. Thus they should be included in early childhood curriculums. It follows that assessment or emergent literacy knowledge and behaviors is integral to quality early childhood reading instruction. Yet, almost all of them remain virtually untapped on statewide assessments of early reading.

To illustrate the relevance of emergent literacy for our examination of issues in early childhood reading assessment, let us examine one aspect of early reading that is important for learning and teaching but ignored in assessment. Following this discussion other aspects of emergent literacy that deserve attention are identified.

One aspect of emergent literacy is emergent storybook reading. Emergent storybook reading might best be thought of as a young child's "reading" of a book. Virtually all young children who are read to at home engage in such behaviors long before they are capable of independent, conventional reading (Sulzby & Teale, 1987). Even a two- or three-year-old will pick up a familiar book, look at the pictures, and proceed to "read" it to a doll, a pet, a parent, or no one in particular. Such behavior is also a common phenomenon among kindergarten children who are read to in the classroom (Martinez & Teale, 1988).

There are a number of different ways that an emergent storybook reading may be done, as Sulzby (1985) has described, and these characteristic ways of reading have developmental properties. For example, at the simplest level, a child may turn the pages of a story book labelling certain items in the pictures ("There's a duck," "Here's the Whatzit") and commenting on the action ("She's running fast! Zoom!") but not weaving a complete story. A more sophisticated emergent reading would be one in which a child uses pictures to recount an oral language-like telling (rather than reading) of the story. At another level the child will sound exactly like she is reading but will attend exclusively to the pictures and often produce language different from what is actually in the text.

For certain types of emergent storybook readings children focus on the print even though they read the book conventionally. For example, when we asked one kindergarten child to read *The Little Red Hen*, he read every word he knew (soft, the, cat, not, I) and skipped all of the other words in the book!

Research has shown that emergent storybook reading behaviors are very important parts of learning to read. These behaviors show what children have learned from interacting with adults in storybook reading situations. Furthermore, they play a key role in helping young children learn about written language (Eller, Pappas, & Brown, 1988; Sulzby, 1985). Such behaviors, therefore, should also be important to early childhood teachers. Instruction should seek to promote emergent storybook readings through a systematic read aloud-program coupled with a well-designed classroom library (Morrow, 1989; Salinger, 1988; Teale & Martinez, 1988; Teale & Sulzby, 1989). Emergent storybook reading however, is not included in any statewide early childhood assessment programs.

This one aspect of early childhood reading illustrates an extremely important point about understanding early childhood reading development and assessing it: one must take into account the child's point of view about what is going on during this period. One must not merely interpret what children do in terms of mature reading conceptions and behaviors. Through the lens of conventional reading, the five-year-old who picks up a book, looks at the pictures, and produces an oral language-like story is doing everything wrong. The implicit implication in current statewide reading assessment programs is that such behavior is not reading, and therefore it is not measured. But when one looks from the child's point of view, it is possible to see that the child is constructing knowledge and strategies for reading. In other words, reading is a thinking process even before it is conventional reading. The more this process is observed and monitored, the more young children can be taught, and they can become fluent, competent readers.

Another area of knowledge that is fundamental to learning to read relates to young children's concepts of the functions and uses of literacy. Heath (1983), Schieffelin and Cochran-Smith (1984), Taylor (1983), Taylor and Dorsey-Gaines (1988), and Teale (1986) have shown that concepts of how reading and writing are used to mediate the activities of everyday life are basic to literacy learning. Understanding that written language functions as a memory aid or a substitute for oral messages provides a basic first step in the long term development of reading skill (Chall, 1983; Teale, 1988a).

Yet another critical aspect of early reading is book handling knowledge and basic concepts about print. Knowing such things as how to hold books (left-to-right, top-to-bottom, front-to-back direction) and the fact that the print, not the pictures is what one actually reads in a book are all important early concepts that children must learn.

The extreme importance of phonemic awareness and a stable concept of word to early reading (Adams, 1989; Juel, 1988; Juel, Griffith & Gough, 1986) also should be considered. Children must be able to segment oral speech into words and, in turn, to segment the words they hear into their constituent sounds in order to accomplish the task of "cracking the code" of the language they are reading. Without phonemic awareness, phonics generalizations about how written language works will never be learned in the way that fluent readers need. Although many assessment programs examine children's knowledge of phonics (sound-symbol correspondences), these programs do not assess oral phonemic awareness. Oral phonemic awareness is, in certain respects, a first step in the process of learning the code of written language. As several researchers have pointed out, without phonemic awareness, children's progress in reading and phonics will most likely be poor.

In summary, our current understandings of several facets of early literacy learning are, in general, not reflected in the reading assessment programs. However, there are some exceptions worthy of careful examination. These include the way that North Carolina's *Communication Skills Assessment for Grades One and Two* (Division of Communication Skills, North Carolina Department of Public Instruction, 1989) uses to assess certain basic concepts about print, and the preliteracy section of the *Metropolitan Readiness Tests* (Nurss & McGauvran, 1985). But overall, most testing programs do not promote instructional activities in reading that recent research suggests should be occurring in the early childhood classroom. Thus, although substantial insight exists into what the content of developmentally appropriate early childhood literacy instruction (and therefore assessment) should be, state assessment methods currently in use tend not to be congruent with that knowledge. We appear to be missing out on assessing major aspects of reading development for these young learners.

What Can Be Done?

An alternative is to change the nature of early reading tests, to bring their content and methods of assessment activities developmentally into line with what is known about effective reading instruction for young children and young children's literacy learning. Assessment programs can go beyond merely collecting report card data. Such data are easy to gather in multiple-choice format and yield some general information of interest to policy makers and the public, but they are virtually useless when making instructional decisions about children. If we are going to devote money and time to statewide early childhood reading assessment programs, we should design those programs to have an effect where it matters most -- in the classroom. Given the high stakes of testing, statewide assessment programs can take a lead in this respect. States can provide valuable leadership in helping local schools focus on early literacy instruction. But early childhood reading assessment programs in the states will have to change considerably in order to perform such a leadership role. I illustrate in the next section of the chapter.

Creating Early Childhood Statewide Reading Assessment Programs

Three changes in statewide early childhood reading assessment can structure testing to be developmentally appropriate and as useful as possible. These changes can be made in purpose, format, and content. All three of the changes are interrelated. I discuss each change individually to consider what actual effects such changes might have on early childhood reading assessment practices.

Change of purpose. The purpose of state assessment programs can be broadened so that a primary goal focuses on providing information useful to the teacher in making instructional decisions about individual children in the classroom. Some might argue that this is already the purpose of such programs, but a closer examination reveals a subtle difference between affecting instruction and helping teachers make day-to-day decisions about developmentally appropriate instruction.

Clearly testing programs affect how teachers teach. However, a recent survey of Texas educators (Teale, 1989b) gives some insight into the perceived nature of these effects. A sample of over 1200 administrators and supervisors who are members of the Texas Elementary Principals and Supervisors Association was asked about the first grade TEAMS test. Although 69% of the administrators agreed or strongly agreed that they were receiving the local support they need to improve TEAMS scores, only 28% of them felt that the emphasis on the test helped teachers to make better instructional decisions. Texas' first grade classroom teachers responded in a similar way. Forty-seven per cent of the over 200 random teachers surveyed said that the reading and writing scores from first grade TEAMS affected their own planning and day-to-day teaching to a great or a considerable extent. Seventy-two per cent agreed that the test had a great or a considerable effect on curriculum and teaching practices in their schools. But 63% said that the effect had been negative or very negative. Furthermore, 90% of the first grade teachers said they would change the practice of assessing first grade children's reading and writing with TEAMS: 45% preferred to replace TEAMS with more developmentally appropriate ways of assessing growth in reading and writing, and 45% preferred to eliminate the test.

Interviews with first grade teachers and kindergarten teachers gave some insight into the reason why the individuals surveyed reacted the way they did. Teachers often said that the test does not give them information about certain facets of learning that they find significant in early literacy development. Teachers also reported that the tests act more as a general survey of achievement for a group of children or a school rather than as a vehicle for helping them instruct particular students. Thus, assessment programs can profitably shift focus more to making assessment an integral part of instruction for individual children in the classroom.

Change of format. Making state assessment programs more closely related to classroom instruction implies that the assessment methods should also change. Formal testing procedures are especially problematic for young children because of the developmental and social characteristics of five- to eight-year-olds. Young children lack experience with test-taking situations and because of the nature of standardized tests, are often easily distracted. Consequently, the relation between test results and actual reading competence can be questionable for children this age.

Formal testing procedures also conflict with the nature of the act of reading for children of this age. Reading is a multifaceted process involving attitudes, knowledge, skill, and self-monitoring, but formal testing procedures are designed to suppress some of these aspects in the attempt to measure a particular feature, or skill. Such a procedure especially affects young children because much of the early learning to read process proceeds from whole to part, with children needing the whole context to be able to display what they know about the parts.

In order to avoid these problems and thereby increase the validity of reading assessment of young children, assessment programs could make increased use of more informal methods. In this way actual acts of reading would become a more fundamental part of the assessment. Such an approach contrasts with the current practice of isolating and testing the various aspects of reading separately. An especially useful way of accomplishing this could be through the use of performance samples. Performance samples are test-like assessment situations in that they center upon predefined aspects of early reading that are to be assessed. However, they are more naturalistic and ecologically valid than testing situations because they yield a record of highly complex behavior on tasks that approximate the reading conditions and resources the students normally encounter in the classroom or other real life settings.

The final format change is that assessment must be conducted more often in order to insure that an accurate picture of young children's reading has been obtained and that the interplay between assessment information and teaching will be a dynamic part of classroom interaction.

Change in content. There is not space here to discuss all of the aspects of early literacy that can profitably be assessed; more detailed information is available in Chittenden & Courtney (1989), Teale, (1988b), and Teale, Hiebert, & Chittenden (1987). In brief, statewide early childhood reading assessment programs should take more of an emergent literacy perspective on content. In so doing, assessment programs can be modified to focus on areas of development like those addressed in the previous section of this chapter.

Putting the Changes into Practice

Two examples of assessment techniques that exemplify the recommended follow. These two examples certainly do not give a complete picture of what statewide early childhood reading assessment programs could be, but they do serve to illustrate the nature of the assessment process that I propose in this chapter. The examples are drawn from research conducted in conjunction with the Chapter One Early Childhood Literacy Program at Albuquerque Public Schools in New Mexico (Teale, 1989a). The instruments were developed in conjunction with Dee Watkins, Linda Harris (currently in Muskogee, OK), and numerous classroom teachers in the Albuquerque Public Schools Chapter One Early Childhood Literacy Program.

The first measure is the "Book Handling and Basic Concepts about Print Task," (BHABCAPT) a procedure which, as its name suggests, assesses children's knowledge of certain book handling conventions (front of the book, the page where one begins reading the book, realization that the print not the pictures is what one reads, direction), concepts about print and words (ability to match speech to print, recognition of what constitutes one and two letters, one and two words), certain conventions of written language (capital letter, punctuation marks), and even certain publishing conventions (concept of title, author, and illustrator). The BHABCAPT is derived from Clay's (1979) *Concepts about Print Test* and the *Book Handling Knowledge Task* of Goodman and Altwerger (1981). The task is conducted on a one-to-one basis and as much as possible, like a regular adult-child storybook reading. The tone of the interaction is kept deliberately informal, more like that of sharing a story than a testing situation. A relatively simply picture storybook, *Ben and the Bear* (Riddell, 1986) is used for the task. Importantly, the book is an authentic piece of children's literature that contains a complete and interesting story. There are predetermined questions that the teacher asks the child during the reading of the story, and thus the task is different from a "real" storybook reading. But the attempt to create a real story reading situation helps make the testing situation ecologically valid because a child's knowledge is assessed within a task that is both purposeful and familiar to the child.

The teacher begins the assessment by handing the child the book upside down and backwards and asks the child to show the front of the book. After the child responds, the teacher suggests that they read the book and asks the child to open it and then to "point to where I start reading." At other points during the administration the teacher does things such as trying to get the child to "follow with your finger as I read" and asking the child about various aspects of directionality and other features of print and conventions noted above.

Results from this task indicate the extent of development in these critical aspects of young children's emergent literacy learning. They also have direct implications for instruction. For instance, the teacher can quickly identify children who have not yet developed the ability to match speech to print and provide them with learning activities like one-to-one storybook reading experiences, shared book experiences (Holdaway, 1979), dictation and rereading of language-experience stories, and *Morning Message* (Crowell, Kawakami, & Wong, 1986) that will help them understand the relationship between oral language and the representation of words in print.

The second procedure is another performance sample of young children's reading, it assesses a child's independent attempts to read stories. For many kindergartners and a substantial number of first graders, this means assessing their emergent storybook readings because they are not yet conventional readers. A description of emergent storybook reading and a discussion of its importance to early childhood literacy development was presented above. Sulzby's (1985) research led to the creation of a classification scale for describing children's emergent readings of favorite storybooks. The scale takes into account such factors as what the child attends to when reading (pictures or print), whether the reading is oral language-like or written language-like, and the ability of the child to produce a story in response to a book. The scale consisted of 11 subcategories that showed the child's increasing sophistication to deal with the text. The 11 categories were too finely specified for use as an assessment instrument, however. Following Sulzby's work on transforming the research scale into an assessment instrument useful in school, we employed a 5-point scale to classify children's reading attempts as follows:

(1) picture governed/no story formed,

(2) picture governed/story formed, oral language-like,

(3) picture governed/story formed: oral and written language mixed,

(4) picture governed/story formed: written language-like, and

(5) print-governed.

In the simplest emergent reading (category 1) children focus on the pictures, label or comment upon the pictures, but do not weave the readings of the separate pages into a story line. With more sophisticated readings represented in categories two, three, and four, children still attend to pictures and create a coherent story across some or all of the pages. The categories advance as the children shift from telling the story (oral language-like) to using the vocabulary, prosody, and structures of written language (written language-like). Readings in which children attend to print are their most sophisticated attempts prior to conventional reading. In this category children may try to sound out words, read only known sight words, give a holistic (and not completely accurate) reading of the text, or even refuse to read based on the realization that they do not "really" know how to read.

By reading a story three or four times to the class and then asking individual children who cannot yet read conventionally to read the story aloud in a one-to-one setting, teachers can determine strategies the children use in their attempts to construct meaning from text. Children's growth over time toward conventional reading can be charted. Thus, a performance sample of children's emergent storybook readings can provide the teachers with useful assessment information. Such an assessment technique fits well with instruction, for it is just these kinds of emergent storybook reading behaviors that repeated storybook readings are intended to develop.

Once the child becomes a conventional reader, performance samples of actual reading can still be gathered for assessment. However, instead of analyzing the readings with an emergent storybook reading scale, miscue analysis of the readings can be performed, along with measuring comprehension of text through a retelling or questioning procedure. This process is like North Carolina's statewide assessment program (Division of Communication Skills, North Carolina Department of Public Instruction, 1989).

These two examples of early childhood reading assessment procedures were presented to illustrate the tenor as well as the content of the approach to statewide assessment advocated in this chapter. Developmentally appropriate assessment of early childhood reading is, at once, informal and rigorous. It enables us to interpret learning from the perspective of the child, and it is theoretically grounded in sound research on written language acquisition. In several respects it looks different from the traditional standardized testing approach to assessment, but it must if assessment of early reading is to move beyond being used as statistical fodder for politicians and the media to becoming legitimate data that teachers, school administrators, and even politicians and the media can use to help children learn to read more effectively and more fluently.

Challenges for Developmentally Appropriate Statewide Reading Assessment Programs

Techniques like the ones illustrated above provide theoretically-based information about students' emerging knowledge and strategies for reading during early childhood. This information relates directly to teaching practices and is therefore directly applicable in the classroom setting. The methods of gathering data are, to a large degree, integral to instruction, and the record keeping techniques would not be too cumbersome. All of these criteria: validity, utility of the information gathered, and ease of use, are important to consider when it comes to evaluating assessment procedures for the classroom. The techniques described here offer promise for satisfying all the criteria. With proper development, their implementation and resultant information about children can have meaning to teachers, to researchers, and to measurement specialists alike, as well as to policy makers and the general public. But major challenges must be met if statewide assessment of early childhood reading development is to move in such directions. I discussed these challenges in another article about information assessment (Teale, 1990) and reiterate them here.

First, there is a need to know more about early childhood literacy learning. We have made tremendous strides recently, but there is still much to be learned about:

- why young children develop literacy,
- what children actually learn,
- how children become literate; and
- when children develop various concepts and strategies.

Especially pressing is the need for information about children from outside the ethnic and cultural mainstream. Developmentally appropriate instruction arises from basic research findings and from carefully conducted classroom research studies. We still have a great deal to discover. That is not to say, however, that development of valid and reliable early childhood literacy assessment procedures must wait upon additional research. Clearly we know enough to take decisive and productive action now; many recent publications in the area indicate that the knowledge base necessary for creating developmentally appropriate measures of early childhood reading exists. But we must continue to insure that assessment procedures reflect what quality research indicates about young children and literacy development.

The second challenge relates to the techniques and instruments themselves. It is clear that there is a paucity of high quality early literacy measures of the type advocated in this chapter. Large scale efforts must be made to develop and field test informal assessment procedures like those discussed in this chapter. For example, Sulzby's (1985) basic research on emergent storybook reading has provided a solid empirical base for the development of an assessment procedure. Such a procedure must now be tested under classroom conditions with representative samples of young children. A high quality instrument sensitive to the range of children being assessed and to the needs of the classroom teacher can be created only with such rigorous development procedures. In other words, it is necessary is to commit money and effort to develop informal measures in a manner analogous to what has been done with large scale standardized tests of early literacy.

In order for the reliability and validity of informal measures to be realized, a third challenge will have to be met. The power of developmentally appropriate early literacy assessment comes from being able to see how young children use their emerging knowledge and skill to accomplish a complex task. Their approach to the task is often not conventional, but almost always rational. The great insight for the teacher comes from understanding what the child has done and why the child has done that. Such a perspective helps in planning instruction. Informal measures help us see early literacy from the child's point of view. It must be recognized, though, that the quality of informal measures is highly dependent upon teacher knowledge. To implement the use of performance samples like emergent storybook readings or the *Book Handling and Basic Concepts about Print Task* or to analyze young children's writing over a six month period, the teacher must know what to look for. To a larger degree than with standardized tests, the instrument for informal measures is the teacher. Therefore, a successful assessment program requires an educational in-service component to help teachers develop their knowledge. I would go so far as to say that efforts to establish informal assessment as a viable tool in the early childhood classroom are doomed to failure without such in-service. States and individual school districts should plan carefully to help teachers understand why such assessment procedures are valuable, what they can learn about children by using them, and how they can use the techniques to interpret the results and apply the information gained. An assessment program that makes extensive use of informal procedures is not as easy to establish as one that relies only upon standardized tests, but the benefits that can be gained are worth the effort.

Finally, there is one challenge that must be met in order for the three previous ones to be attempted. An integral part of the development of an informal assessment program must involve political considerations. If informal assessment is to be implemented in our early childhood classrooms, it must be legitimized. Informal assessment carries considerably less weight in the school decision making process than standardized measures. Informal assessment is also often viewed with suspicion by measurement personnel in school districts. Standardized tests are usually the only measure of accountability or effectiveness that public of policy makers know or use. Despite the fact that teachers do not find the results from standardized tests very useful in daily classroom planning and instruction, teachers tend to hold standardized tests in considerable esteem. There are no doubt numerous reasons why informal measures do not occupy the status of formal measures. But a critical goal for states is to help convince everyone that informal measures can be just as accurate, reliable, valid, and even more useful for instructional purposes.

For this to happen we must, of course, meet the challenges of the three previous points mentioned. Thus, our challenges are inextricably intertwined. They must be seen as part of an overall movement that is at once an issue of instruction, as well as an issue of measurement, and an issue of politics.

Conclusions

Statewide assessment of early childhood reading is conducted in most states with multiple-choice tests that measure aspects of reading such as sight word recognition, phonics, main idea, sequencing events, word structure, and comprehension of short passages of text in a manner similar to the way in which these skills are tested with older children in the elementary and middle schools. Such a practice is often problematic because the methods of assessment fail to consider social and developmental characteristics of young children and because the tests do not assess many important aspects of children's early reading knowledge and behavior.

A backlash against intensive standardized testing of young children has caused some states to eliminate reading tests in kindergarten, first grade, and even second grade. This chapter has tried to show that there is an even better alternative. Statewide assessment of early childhood reading can be done in a way that is developmentally appropriate, that pays attention to the knowledge and behaviors which current research shows are significant to early reading development, and that relates closely to classroom instruction. Furthermore, states can take a proactive stance by implementing such assessment programs because a theoretically sound and technically valid and reliable assessment program will actually serve to improve instruction in the classroom.

Such an approach to statewide assessment would require major changes in the nature of early childhood reading assessment. Assessment must become more informal and utilize one-to-one procedures and performance samples much more extensively. Assessment must also be done on a more frequent basis so that it becomes part of a process of true diagnostic teaching. Finally, new procedures that assess the heretofore unmeasured, yet significant, aspects of early reading development must be implemented as part of any overall assessment program.

Such a proposal brings with it several challenges for research, for instrument development, for teacher education, and even for educational policy. It will not be simple, but it is certainly possible to move statewide assessment programs of early childhood reading in more developmentally appropriate directions in the coming decade. In this way assessment could make its greatest contribution to better reading instruction for young children.

References

Adams, M. (February, 1989). *Phonics and beginning reading instruction.* (Final Report). Champaign, IL: Reading Research and Education Center.

Chall, J. (1983). *Stages of reading development.* New York: McGraw-Hill.

Chittenden, E.A., & Courtney, R. (1989). *Assessment of young children's reading: Documentation as an alternative to testing.* In D. Strickland & L.M. Morrow (Eds.), *Emerging literacy: Young children learn to* *read and write*. (pp. 107-120). Newark, DE: International Reading Association.

Clay, M.M. (1979). *The early detection of reading difficulties* (3rd ed.). Auckland, New Zealand: Heinemann.

Crowell, D.C., Kawakami, A.J., & Wong, J.L. (1986). Emerging literacy: Reading-writing experiences in a kindergarten classroom. *The Reading Teacher* ,40, 144-149.

Darling-Hammond, L. & Wise, A.E. (1985). Beyond standardization: State standards and school improvement. *Elementary School Journal*, 85, 315-336.

Division of Communication Skills, North Carolina Department of Public Instruction. (1989). *Communication skills assessment for grades one and two*. Raleigh, NC: North Carolina Department of Public Instruction.

Eller, R.G., Pappas, C.C., & Brown, E. (1988). The lexical development of kindergartners: Learning from written context. *Journal of Reading Behavior*, 20, 5-24.

Goodman, Y.M. & Altwerger, B. (1981). *Print awareness in pre-school children: A Working paper*. A study of the development of literacy in preschool children, Occasional Paper No. 4. Tucson, AZ: Program in Language and Literacy, University of Arizona.

Heath, S.B. (1983) *Ways with words: Language, life and work in communities and classrooms*. Cambridge, U.K.: Cambridge University Press.

Holdaway, D. (1979). *Foundations of literacy*. Auckland, New Zealand: Ashton Scholastic.

Juel, C. (1988, April). *Learning to read and write: A longitudinal study of fifty-four children from first through fourth grade*. Paper presented at the Meeting of the American Educational Research Association, New Orleans.

Juel, C., Griffith, P.L., & Gough, P.B. (1986). Acquisition of literacy: A longitudinal study of children in first and second grade. *Journal of Educational Psychology*, 78, 243-255.

Madaus, G.F. (1988). "The influence of testing on the curriculum." In L.N. Tanner (Ed.), Critical issues in curriculum: 87th yearbook of the National Society for the STudy of Education. Chicago: University of Chicago Press.

Martinez, M.G. & Teale, W.H. (1988). Reading in a kindergarten classroom library. *The Reading Teacher*, 41, 568-573.

McNeil, L.M. (1988). "Contradictions of Control, Part 3: Contradictions of Reform." *Phi Delta Kappan*, 69, 478-485.

Miller, G.A. (1977). *Spontaneous apprentices: Children and language*. New York: Seabury Press.

Morrow, L.M. (1989). *Literacy development in the early years*. Englewood Cliffs, NJ: Prentice-Hall.

National Association for The Education of Young Children (1988). NAEYC position statement on standardized testing of young children, 3 through 8 years of age. Young Children, 43, 42-47.

National Association of Early Childhood Specialists in State Departments of Education (1987). *Unacceptable trends in kindergarten entry and placement: A position statement*. Lincoln, NE: National Association of Early Childhood Specialists in State Departments of Education.

National Association of State Boards of Education (1989). *Right from the start: The report of the NASBE task force on early childhood education*. Alexandria, VA: National Association of State Boards of Education.

Nurss, J.R. & McGauvran, M.E. (1986). *Metropolitan readiness tests*. Orlando, FL: Harcourt, Brace, Jovanovich.

Riddell, C. (1986). *Ben and the bear*. New York: Harper and Row.

Salinger, T. (1988). *Language arts and literacy for young children*. Columbus, OH: Merrill.

Salmon-Cox, L. (1982, September). *MAP math: End of year one report*. Pittsburgh, PA: Learning Research and Development Center.

Salmon-Cox, L. (1984, September). *MAP reading: End of year one report*. Pittsburgh: Learning Research and Development Center.

Schieffelin, B. & Cochran-Smith, M. (1984). *Learning to read culturally: Literacy before schooling*. In H. Goelman, A. Oberg, & F. Smith (Eds.), Awakening to literacy (pp. 3-23). Exeter, NJ: Heinemann Educational Books.

Shepard, L.A. (1989). Why we need better assessments. *Educational Leadership*, 46, 4-9.

Stiggins, R.J. (1985). Improving assessment where it means the most: In the classroom. *Educational Leadership*, 43, 69-74.

Strickland, D. & Morrow, L.M. (1989). *Emerging literacy: Young children learn to read and write*. Newark, DE: International Reading Association.

Sulzby, E. (1985). Children's emergent reading of favorite storybooks: A developmental study. *Reading Research Quarterly*, 20, 458-481.

Sulzby, E. & Teale, W.H. (1987). *Young children's storybook reading: Longitudinal study of parent-child interaction and* *children's independent functioning*. Final Report to The Spencer Foundation. Ann Arbor, MI: The University of Michigan.

Sulzby, E. & Teale, W.H. (in press). *Emergent literacy*. In P.D. Pearson, R. Barr, M.L. Kamil, & P. Mosenthal (Eds.) Handbook of reading research (2nd ed.). New York: Longman.

Taylor, D. (1983). *Family literacy: Young children learning to read and write*. Exeter, NJ: Heinemann.

Taylor, D. & Dorsey-Gaines, C. (1988). *Growing up literate: Learning from inner-city families*. Portsmouth, NJ: Heinemann.

Teale, W.H. (1986). *Home background and young children's literacy development*. In W.H. Teale & E. Sulzby (Eds.), Emergent literacy: Writing and reading (pp. 173-206). Norwood, NJ: Ablex Publishing Corporation.

Teale, W.H. (1988a, November). *Becoming literate: The first step in learning to read and write*. Paper presented at the 22nd Annual California Reading Association Conference, San Diego, CA:

Teale, W.H. (1988b). Developmentally appropriate assessment of reading and writing in the early childhood classroom. *Elementary School Journal*, 89, 173-183.

Teale, W.H. (1989a). [Albuquerque Public Schools early literacy assessment project.] Unpublished raw data.

Teale, W.H. (1989b). *Early literacy assessment in Texas: Recommendations for change*. Texas Reading Report, 11, pp. 1, 9, 11.

Teale, W.H. (1990). *The promise and challenge of early literacy assessment*. In L.M. Morrow & J.K. Smith (Eds.), Assessment in early literacy instruction (pp. 45-61). Englewood Cliffs, NJ: Prentice-Hall.

Teale, W.H., Hiebert, E.H., & Chittenden, E.A. (1987). Assessing young children's literacy development. *The Reading Teacher*, 40, 772-777.

Teale, W.H. & Martinez, M.O. (1988). Getting on the right road to reading: Bringing children and books together in the classroom. *Young Children*, 41, 10-15.

Teale, W.H. & Sulzby, E. (Eds.). (1986). *Emergent literacy: Writing and reading*. Norwood, NJ: Ablex.

Teale, W.H. & Sulzby, E. (1989). *Emergent literacy: New perspectives on young children's reading and writing development*. In D. Strickland & L.M. Morrow (Eds.), Emerging literacy: Young children learn to read and write (pp. 1-15). Newark, DE: International Reading Association.

Valencia, S.W., Pearson, P.D., Peters, C.W., & Wixson, K.K. (1989). Theory and practice in statewide reading assessment: Closing the gap. *Educational Leadership*. 46, 57-63.

Chapter 4

The Role of Teacher-Based Information in Statewide Assessments of Literacy Learning

The Role of Teacher-Based Information in Statewide Assessments of Literacy Learning

Elfrieda H. Hiebert
University of Colorado, Boulder

While a previous generation worked hard to develop specific objectives and criterion-referenced items to assess those objectives, the current generation has realized that a skills-driven model of curriculum, instruction, and assessment does not add up to the whole. An era of test-driven curriculum has produced high performances on multiple-choice tests but low performances on tasks that require synthesis (Applebee, Langer, & Mullis, 1989). While the relationship between assessment and curriculum should, by definition, be inextricably interwoven, difficulties arise when curriculum is matched perfectly with the content and formats of multiple-choice tests (Shepard, 1988). The resulting problems of this match between curriculum and multiple-choice tests have led educators to examine more closely the goals of schooling and the manner in which current assessments reflect critical goals. This effort has led several states to explore alternative means of capturing the critical goals of literacy.

One solution has been to create better paper-and-pencil tests. The efforts of Illinois (Valencia & Pearson, 1987) and Michigan (Wixson, Peters, Weber, & Roeber, 1987) show that tests can be developed that better represent a view of reading as the construction of meaning. Illustrative of these efforts, the Illinois test uses longer passages, increases the demand for student reasoning by using multiple multiple-choice formats and higher-level questions, and assesses students' prior knowledge about the topic and their application of reading strategies.

States such as Vermont represent a second solution which is to integrate student work portfolios into state assessments (Brewer, 1989). These student portfolios have their precursors in the writing samples that many states and districts have used for a number of years (Chapman, 1988; Meredith-Dabney, 1988; Vickers, 1988). The typical mode of district and state assessments of writing has been to obtain samples of students' writing, with topic and genre held constant, and to have someone other than the classroom teacher analyze samples. Sometimes these third parties consist of teachers but usually not students' classroom teachers. In a similar vein, the aim of portfolios under investigation by state departments of education place fairly tight strictures on portfolios. While several projects indicate that portfolios can be used in inventive ways (Archbald & Newman, 1988; Wolfe, 1989), the use of student portfolios does not necessarily involve more refined assessment strategies on the part of teachers nor does it necessarily draw on what teachers know. The integration of information gathered by teachers in their classrooms into district and state assessments is the concern of this chapter. This possibility does not rule out the use of portfolios. As Valencia, McGinley, and Pearson (in press) suggest, portfolios can have different components, including information specified by policy-makers, that selected by teachers, and that selected by students in collaboration with teachers. Furthermore, this aim of integrating teacher-based information is viewed in conjunction with other solutions to the assessment problem, not in competition with them.

The first part of this paper provides a rationale for inclusion of teacher-based information in state-wide assessments, the second part illustrates the forms that teacher-based assessment can take, and the concluding part of the paper presents several issues that require changes in perspective for teacher-based assessment to become an integral part of state and district assessments.

Some states have put into place performance-based assessments. Performance-based assessment does not necessarily involve teachers in different assessment practices, since outsiders may come into classrooms to assess children or children may go to central school or district sites to participate in these assessments as in the state of New York's science assessment (Reynolds, 1989). However, performance-based assessments can be designed so that teachers are responsible for gathering and/or analyzing information. Whenever possible, these cases will be used to illustrate the manner in which states and districts can support teacher-based assessment. Since American schools fall short on high-level literacy skills rather than low-level ones, examples will focus on a basic dimension of literacy that is infrequently assessed in paper-and-pencil tests -- students' abilities to interpret information critically.

Why Teacher-Based Assessment?

While this list is not exhaustive, three reasons for integrating data from teacher-based assessment into state and district assessment programs can readily be identified.

Information on the "higher literacies"

As Brown (1989) discusses, the "higher literacies" -- abilities such as establishing the bias of a writer or speaker and synthesizing information from several sources -- characterize the literate person in the information age. As teachers interact with students in numerous contexts over a school year, they have the opportunity to observe students' application of the higher literacies. They hear students' responses to questions, as well as their questions. They read students' compositions at different points in a school year on self-selected and teacher-directed topics and across subject areas. Teachers' access to information is much more extensive and encompasses many more contexts than the standardized testing situation which captures children's responses in one setting and at one point in time. While standardized testing may be said to be more reliable than teachers' judgments, teachers may well make up for this through the authenticity of the situations in which they see children and the extensiveness of the data.

Building on teachers' instructional expertise

Teachers' sharing and gathering of information about student accomplishments can have the added benefit of improving instruction. The close match between assessment and curriculum is rightfully criticized when multiple-choice tests are the source of curriculum, as is the case in measurement-driven instruction (Popham, Cruse, Rankin, Sandifer, & Williams, 1985). However, when the goals of schooling are defined more broadly and when the measures that assess attainment of these goals allow a range of tasks and response formats, the match between assessment and instructional practices should be close.

In this latter scenario, making explicit the goals of schooling and the means of assessing these goals can assist teachers in the quality of instruction they provide. Assessment of students' ability to detect the bias of an author might be assessed with a set of newspaper columns. Examples of such assessment activities and evidence of what constitutes detection of authors' biases can help teachers provide instruction on this critical goal of schooling. Some school districts and state departments of education support such integration by providing examples of assessment and instructional activities alongside objectives in their curriculum guides.

A close link between instruction and assessment furthers the goals of literacy to the degree that the assessment activities reflect what is known about proficient literacy use. When assessment practices capture trivial goals, instruction that mimics assessment instruments may do little more than create smart test-takers. When assessment tools provide information on the critical goals of literacy, teachers' familiarity with state and district assessment activities and coordination of their instructional practices with these activities should not be viewed as a surreptitious act but as part of appropriate instruction (Shulman, 1988).

Increasing teachers' involvement in the educational process

Teachers are the ones who impact students but yet they rarely have any say or input into policy-making. In turn, policy-makers rarely see classroom life in action. The relationship between policy-makers and teachers is often an adversarial one. Policy-makers do not trust teachers' judgments; teachers do not believe policy-makers' mandates to be valid relative to their contexts. Even small steps in integrating teachers' information about their students into decision-making beyond the classroom can be expected to go a long way. When teachers' information is used beyond the classroom, teachers document dimensions of classroom life frequently left unarticulated (see Amarel & Chittenden, 1982). This process can have the end result of providing information to policy-makers on critical dimensions of literacy that are typically uncaptured. In addition, teachers' ownership of the educational process can be expected to increase.

A Model of Teacher-Based Assessment

Various suggestions have been made about teachers as reflective practitioners who base decisions on information that they have gathered on students (Clark & Peterson, 1986). Teachers require a problem-solving stance toward students' learning and classroom events, similar to that of a chemist or an architect as they work to solve a problem. Problem framer and solver aptly describes the role of teachers (Calfee & Hiebert, 1988). Instruction, curriculum, and assessment are interwoven as teachers set goals, gather data, and make decisions.

Goal Setting

The establishment of goals underlies the entire educational enterprise. The relationship of teachers in establishing literacy goals is a difficult one to define. While teachers need to clearly articulate their goals, the goals of schooling also reflect the larger community. The balance between teachers' translation of goals to their unique settings and the identification of goals by the larger community is often not addressed in districts and states. Teachers enter the profession with visions, beliefs, expectations, and perceptions. Often, these ideas run counter to efforts of state departments of education in operationalizing goals. A state department of education may be well-intentioned in its identification of comprehension as a priority but its translation of this goal to specific items on a competency test may run counter to teachers' broader interpretation of comprehension. One step toward a common sharing of goals is for school faculties to discuss the translation of district or state goals in their schools.

A shared vision of a literate individual is at the heart of these discussions. Descriptions of such visions exist in several places. *Becoming a Nation of Readers* (Anderson, Hiebert, Scott, & Wilkinson, 1985), for example, described a view of readers as constructive, strategic, fluent, and motivated. Calfee's (1988) vision includes goals related to comprehension of expository and narrative text, decoding, and vocabulary.

A general vision of a literate individual is only the first step. A shared image of this individual at different points in development is critical for a faculty of teachers. Most teachers identify their primary goal to be the creation of readers who enjoy reading and read extensively. Teachers in a school can benefit greatly from describing the manifestation of this goal in a first grader versus a sixth grader. The sixth grader, for example, might be expected to be much more involved with informational material, while the first grader's interest might be displayed in read-aloud contexts.

Gathering data

A perusal of textbooks on reading pedagogy produces a number of techniques which teachers are advised to use for collecting information on their students. Informal reading inventories, checklists, surveys, teacher-made tests, miscue analysis, observational schemes, kidwatching, performance samples, and portfolios are among these. These presentations frequently fail to make the unique functions of different techniques clear. A contrasting view is a framework of the processes of teacher-based assessment. From this perspective, teachers can gain four types of data about their students that differ from that of typical paper-and-pencil, standardized tests.

Three of the processes are distinct from one another in the activity that is implied on the part of the teacher: observing, questioning, and examining student work samples. All of these provide information on different dimensions of student learning processes and products. There is redundancy, of course, in that facets of the same proficiency can be examined by the three processes. If, for example, students' facility in writing expository text were of interest, a teacher might observe the kind of support students receive from one another during writing. Next, the teacher might examine students' compositions to determine facility with various text structures. These compositions could become the basis for an interview in which the teacher questions students about their use of text structures. Each teacher assessment process sheds light on different dimensions of student processes and products.

A fourth form of teacher-based assessment -- guiding students in self-assessment -- varies somewhat from the other three processes in that instruction is more directly involved. This activity overlaps with the other assessment processes since teachers might assess students' facility in self-assessment through observing, questioning, and sampling evidence of self-assessment. Guiding students in self-assessment is included here because this dimension of teacher-based assessment may, ultimately, be the most important. Students' ability to accurately evaluate strengths and weaknesses is a goal of literacy instruction that is often overlooked.

Observing. When teachers are asked about the forms of assessment that determine their instructional actions and lanes, they typically cite their observations first, with sources of information such as standardized tests falling far behind (Dorr-Bremme & Herman, 1986; Salmon-Cox, 1981). Teachers are in a continual process of observing their students. While they may see these observations as critical sources of information, teachers' observations can be ill-formed (Gil, Polin, Visonhaler, & Van Roekel, 1980). What might initially appear to be capriciousness in teachers' evaluations can be traced to a minimal, and often nonexistent, foundation. Most teacher education programs treat the topic of teacher-based assessment superficially at best (Schafer & Lissitz, 1987). However, even a little guidance goes a long way. A training session as short as a month can increase the consistency of teachers' evaluation of data considerably (Gil et al., 1980).

Observational data should be grounded in a vision of the critical dimensions of literacy at particular levels. Teachers do not have to wait for particular events to occur so that they can observe their students; instructional contexts can be created that allow the gathering of particular information. To obtain information on students' abilities to analyze authors' points of view, for example, a teacher might set up discussions in which students talk about points of view in familiar events such as the season's popular television shows.

A benefit of observational data is that information can be gained on students' behaviors in everyday situations. Many students, but especially those whose backgrounds are unlike academic environments, respond negatively in evaluative contexts (Hill, 1984; Mosenthal & Na, 1980). Students' interactions in groups as compared to individual settings, such as a one-to-one discussion with the teacher, can be documented.

Gains in authenticity do not have be at the sacrifice of reliability. As illustrated in the Gil et al. study, guidance and practice increases the consistency of observations within and across teachers. Opportunities for teachers to direct their attention to critical dimensions of literacy are a first step in the process of gaining trustworthy information. Observations can be aided considerably when teachers keep records. One suggestion is that teachers take notes of particular activities or students much like an ethnographer might (Marzano, Hagerty, Valencia, & DiStefano, 1987). Checklists can also be helpful in documenting observations. For example, a checklist that identifies processes of efficacious literature response groups can assist teachers in studying students' learning and in facilitating groups.

Questioning. Settings where teachers and students discuss, either around a systematic set of questions or otherwise, provide another means of gathering data. In this case, the emphasis is on oral expression -- which is itself a critical proficiency and one in which children are typically more facile than in writing. While most formal testing is done with paper-and-pencil tasks, people most frequently map courses of action in interchanges of ideas between neighbors, family, and co-workers. In corporate settings, courses of action are often established in a perpetual round of meetings rather than in solitary, written contexts. Assessment of students' understandings and applications of strategies in the contexts of teacher-student and student-student interaction clearly is important. Questioning permits in-depth assessment of students' interpretations, unencumbered by their ability to write.

Like other dimensions of assessment, teachers' questions need to be guided by some theoretical perspective. Recent work on story structure, which has been presented to teachers in a variety of materials, provides an excellent means for guiding teachers' questioning. Students' failure to grasp the plot of a story, for example, is useful information to teachers. The framework of story structure also makes it easy for teachers to document children's responses. A simple form can be used to summarize students' comprehension at different points in the school year and with different genres, such as mysteries and science fiction.

Sampling. Of all alternative assessment techniques, portfolios or collections of student work have most captured the interest of educators. The original use of the term portfolio came from the collections of artists and architects who keep samples of the best of their work. In current usage, portfolios consist of examples of students' work over time and in particular tasks such as an essay, a narrative, and a persuasive piece. While the idea of portfolios as highlighting students' "best work" has not been the typical interpretation in school settings, the concept of portfolios is serving to restructure assessment activities of school districts and state departments of education. Obviously, teachers can sample student work without a portfolio system. Students' comments about point-of-view in narrative passages might be compared to their analyses of point-of-view in expository passages.

Samples of writing can be obtained much more readily for portfolios than samples of students' reading. As a consequence, the shifts in assessment have been much more dramatic for writing than for reading. Many states and districts evaluate actual samples of students' compositions for their writing assessments, in addition to or as a substitute for standardized tests which typically emphasize mechanics.

Efforts at performance-based reading assessment are beset with many more difficulties than those with writing. In reading, the reform of assessment has been manifest most clearly by the improvement of multiple-choice, standardized tests. Existing performance-based assessment efforts are relying on the ease of gaining writing samples by assessing students' reading through written responses. The ability to express one's self in writing is obviously a critical dimension of sharing interpretations from reading but an over-reliance on written formats disregards other critical dimensions of reading. At the beginning stages of reading, for example, application of different cuing systems may become most apparent in an oral reading and retelling.

While a long history exists on performance-based reading assessment in the form of the informal reading inventory, dating back at least to Gray (1920), this activity has never captured the interest of policy-makers (Johnston, 1984). Even Goodman's (1968) miscue analysis which reconceptualized oral reading and retelling in a psycholinguistic framework has failed to generate greater use of analysis of oral reading and retelling samples. This approach appears to be too cumbersome and, unlike writing samples which can be gathered at a central place and quickly scored, requires either on-the-spot scoring or tedious transcription of audiotapes. The assumption of "the same amount of data for all children" acts against the use of oral reading samples. Another reason may well be a distrust of teacher integrity in doing on-the-spot scoring (Johnston, 1984).

Several recent efforts are worthy of review because they use writing in inventive but not overly-taxing ways. The efforts of the National Foundation for Educational Research (NFER) in England and Wales are especially noteworthy since they illustrate large-scale use of innovative assessments (Assessment of Performance Unit, 1987). Over the past decade, the NFER has used authentic passages of three types (works of literature, works of reference, and everyday reading materials such as brochures, bus schedules). Children read a passage of some length and substance (e.g., a complete brochure in the case of everyday reading materials or an intact piece of literature) and write responses to questions about the material. Questions require a range of factual and interpretive application in the form of writing. For example, eleven year olds were asked to write about an amusing part of the story *Nothing to be Afraid of*, a passage with comic elements. Questions about a brochure required students to apply information to the needs of a particular family. Not only did students need to retrieve information that was explicitly stated in the brochure but they also needed to use information about family members (e.g., Jane was an active sportsman; Michael was not) in interpreting information in the brochure (e.g., "Would the sports and entertainments offered [in Warminster on Sea] appeal more to someone who participated or to someone who just wanted to watch?").

With new mandates for national assessment in England and Wales, the NFER is moving on to another stage (Burstall, 1989). They are in the process of pilot testing "integrated tasks" that include cross-curricular components and occur in instructional contexts. A sample task occurs in the context of a group of students with their teacher. Children have a booklet pertaining to an experiment about the characteristics of different materials. Students make predictions about the durability of a set of materials that includes paper and aluminum foil in relation to different actions (e.g., placing the materials in water, rubbing them). The teacher engages students in a discussion about predictions regarding the action, after which children write down predictions individually. Children then execute the experiment, recording the data from the experiment in their individual booklets. The group finally discusses conclusions that can be drawn from the experiment, followed by children individually writing down their conclusions. While the processes of predicting and observing are part of reading, these integrated tasks are not as direct an assessment of reading strategies as the earlier NFER assessments. However, these efforts do illustrate a commitment to placing assessment, even that to be used at the national policy-making level, within the contexts of classroom instruction.

The efforts of Massachusetts illustrate some current efforts in the United States in which assessment tasks more closely mirror those of effective reading instruction (Massachusetts Department of Education, 1987). In a typical task at the fourth-grade level, students are asked to predict the contents of a passage from the beginning sentences of an article. A typical item reads, "A newspaper article is entitled, 'Lake Champlain's Monster -- Fact or Fiction?' The first two sentences of the article are:

'Believers say the warm waters of summer bring the monster to the surface. Others say that the monster is just the creation of jokesters.' Describe the kind of information you expect the rest of the article to contain." This task assesses a skill that is an earmark of the proficient reader -- the ability to activate expectations and prior knowledge relative to a topic. While a review failed to locate any state efforts to assess reading comprehension through oral communication, one project was located that assesses group and individual processes in discussion contexts -- an assessment of high school mathematics (Rindone, 1989). The preliminary conceptualization of this project that is part of Connecticut's new assessment program involves teachers watching group processes and giving on-the-spot ratings for groups and individuals within groups. Some of the criteria focus on communication processes, while others focus on drawing accurate conclusions from the data. In addition to analyses of students' group and individual communication in the group setting, students' final reports are evaluated individually. Journals are included where students comment on their processes and products.

It is important to note that an on-the-spot rating system of student oral language has begun in high school mathematics and that efforts to integrate oral responses in literacy are less evident. This may reflect more respect for high school content area teachers' expertise, although the Connecticut Assessment will ask mathematics teachers to focus heavily on group communication processes, a skill that typically is not associated with the expertise of high school mathematics teachers.

While writing is obviously much easier to obtain in a portfolio than samples of students' reading or speaking, advancements in technology mean that students' oral reading and discussions can be captured on videotapes and audiotapes. A record of a class debate on the biases of newspaper articles could be an important scenario for teachers and their students to study. Events that in generations past went unrecorded can be documented and reflected upon by different groups and on many occasions.

Another possibility is evidenced in the assessment practices of the Coalition of Essential Schools (Wiggins, 1989) which emphasizes demonstrations or exhibits of student learning. An analogue for these demonstrations in learning outside the school would be the demonstrations required of Boy Scouts to obtain a particular badge. In an elementary classroom, demonstrations might take the form of exhibits similar to those at science fairs. For example, the books that students have written over the course of a school year could be displayed at a book fair or a classroom might be set up as museums often are, with students' reports mounted beside artifacts related to their topics.

Unlike observations, samples of student work can be reviewed again and again for different purposes. Students and teacher can, separately and together, reflect on progress. Different groups beyond the classroom can also independently evaluate samples and come up with unbiased conclusions.

Guiding students in self-assessment. In most classrooms, the teacher is the judge of students' accomplishments. Students have few, if any, opportunities to evaluate their progress, much less create projects and establish the means of completing them. Teachers are evaluated by external mandates and they, in turn, create a system that is externally-driven. Such an external system works against the self-monitoring and regulation that marks effective completion of projects in domains beyond the school. In most arenas, effective participation depends on one's ability to establish goals and ways of achieving these goals and to monitor progress toward these goals. Such processes need to be a built-in part of school activities.

Project Zero stands in sharp contrast to the externally-driven assessment of typical classrooms (Wolfe, 1989). The portfolios developed in Project Zero include two elements: items such as student compositions and their reflections on products. These reflections take the form of diaries or journals in which students compare and contrast their work. Furthermore, students decide what will be included in their portfolios, with the teacher privy to the decision-making process but students making the final decisions. As this illustration shows, portfolios by no means should be viewed as a new form of externally-mandated assessment. Teachers, and their students, can be inventive with portfolios in ways that further self-assessment.

Decision-making

The aim of increased attention to teacher-based assessment is to extend teachers' use of information in refining curriculum goals and instructional processes. If teacher-based assessment is another task added to the already heavy load of teachers, the purpose has not been realized. Assessment should be viewed by teachers as part and parcel of their programs. For some teachers, such a stance may be novel as illustrated by the responses of a group of teachers. When asked about the success of a recent move to literature-based reading instruction, teachers said that their students were having "more fun" but they were hard-pressed to document an increase in students' enjoyment and involvement in reading. Data were necessary, however, for a school board that continued to show a concern for results. Teachers had been unaware of evidence that was readily apparent to an observer, such as the level of writing and amount of involvement in the annual author's fair held by the school. As teachers began to take a new view of assessment, they saw numerous means of providing the school board with evidence of students' participation as avid readers and writers.

Teacher-based assessment can best be seen as a cyclic process, with new questions raised as teachers assess. In working toward students' critical listening and reading, a teacher discovered that students raised questions about incongruities and events on television shows and movies. Since students' viewing was limited to "narrative" and not informational television shows, the next step was to determine whether they raised questions about bias and point of view in television news.

Ultimately, a goal is for teacher-based assessment to enter into district and state decision-making. Such use is predicated on a changing of perceptions, as is developed next.

Next steps for integration of teacher-based assessment into decision-making in and beyond the classroom

A Catch-22 exists with regard to teacher-based assessment. Teachers do not document information because no one asks them to share this information. Administrators and policy-makers claim that teachers are not systematic about their observations and evaluations of student work. When teachers are asked to provide administrators with data, their documentation becomes more extensive (Amarel & Chittenden, 1982).

Efforts to further teacher-based assessment need to be two-pronged. Before describing these two prongs, it should be recognized that both prongs require at least a modicum of resources. Resources depend on commitments from administrators. Even so, teachers do not have to feel that the matter is out of their hands. In one district, school board members received an unsolicited report from one school on the numbers of books that students had authored and samples of this work. Reeducating the public about the critical goals of literacy also occurs through newspaper articles and, at a very local level, the integration of reflective teacher data in parent-teacher conferences.

The two prongs related to teacher-based assessment have to do with fundamental changes in conceptions about assessment that underlie district and state mandates and, second, opportunities for teachers to observe, document, and analyze children's learning and instructional opportunities.

Changes in fundamental concepts about assessment

Some deeply-held conceptions about assessment need to change for teacher-based assessment to become part of district and state assessment processes. One pervasive assumption that underlies American school evaluation and that limits what is possible in evaluation is that the same amount of data must be gathered on all individuals. This has a limiting effect on what can be assessed. In a state like California, for example, the presence of approximately 300,000 youngsters at a grade level sets limitations on the kind of information measures that can be gathered. When a decision was made several years ago to sample students' writing, the view was that a sample of writing needed to be evaluated for every child at a particular grade level. Thus, a massive financial commitment was made to sample a composition from every eighth-grade child in the state. Even with a matrix sampling procedure in which the genres on which students wrote were varied, the evidence from students was limited and, given the influence of such features as topic on amount and quality of writing (Scardamalia and

Beireter, 1986), the meaning of the evaluations, once they had been gained, was unclear.

An alternative is to collect in-depth information on a subset of students. If certification is the issue, as is the case with the certification of teachers (Shulman, 1988), gathering comprehensive data on every individual is critical. However, if patterns of accomplishment are of interest and in-depth measures allow studying goals that are otherwise overlooked, the situation is entirely different and a sampling procedure might well be appropriate for some dimensions of literacy. For other dimensions, all students may be assessed with more easily scorable measures. A view of "the same data for all" does not drive those studying political or consumer views. The techniques of representative sampling have been perfected by American pollsters. If educators did not have to meet the criterion of similar data for all constituents all of the time, more in-depth information could be gained about critical dimensions of literacy.

Opportunities for teachers to study children and instruction

The skills of teachers in studying children's progress as well as their instruction require basic development. Opportunities for such study depend on fundamental changes in teacher education and staff development. From every indication (see, e.g., Dorr-Bremme & Herman, 1986; Schafer & Lissitz, 1987), teachers receive very little guidance in assessing students in either preservice training or later staff development. If teacher education is a set of hurried "how-tos" that fail to engage teachers in reflection, it should come as no surprise that teachers have not developed a stance of goal-setting and decision-making. Dramatic changes are required in teacher education to provide the experiences that create such a stance. A structure for these changes is present in Berliner's (1987) proposal of the laboratory in teacher education. One component consists of field-based experiences where teachers-to-be observe, interview, sample student work, and act on information in actual classroom settings. Berliner, however, uses the term laboratory in another manner, similar to the way in which it might be used in chemistry or biology where students conduct experiments. The experiments for teachers-to-be use videotapes, audiotapes, and transcripts of classroom events that require them to reflect on and apply information. Opportunities such as these allow teachers-to-be to mull over information, detect patterns, and analyze instruction and student learning. Information in the form of transcripts, audiotapes, and videotapes is, in the long run, more accessible and capable of reflection than are on-the-spot classroom observations. Videotapes permit analysis and reanalysis -- truly the stance that is desired of the "kid-watcher," as Goodman (1985) labels the role of a teacher. While few teacher education programs currently provide the opportunities that Berliner describes, videotapes of classrooms with accompanying materials to be used for analysis and reflection are beginning to appear (see, e.g, Anderson, Au, Borko, Guthrie, Hiebert, & Mason, 1987).

Changes are also needed in the school context for teachers to act as data-gatherers and decision-makers. The task of supporting teacher-based assessment is a very different one in schools than the task which confronts preservice training efforts in universities. Preservice training should develop a stance in teachers-to-be toward data-gathering and decision-making and the basic skills to perform those roles. Schools and districts need to provide collegial environments in which teachers have opportunities to interact with one another about the information that they have gathered.

Conclusion

Forms of assessment other than standardized, multiple-choice tests are clearly needed. When information from alternative measures is considered, it becomes clear that teachers are important elements in statewide assessment programs. Teachers' information represents the literacy tasks that individuals confront in life more authentically than the multiple-choice items of standardized tests. Instructional processes such as observing, questioning, and sampling student work provide a wealth of information about students' literacy abilities in day-to-day settings.

The breadth of teachers' information and the sheer numbers of teachers within a state may make the task seem formidable and even impossible. National efforts in England and Wales, however, show that careful design can combine instruction and assessment. State efforts like those of Connecticut and Vermont illustrate that new assessment instruments can be designed that integrate, to a greater extent, teachers' knowledge and skills.

Teachers are an indispensable source of information about their students' literacy accomplishments. While efforts that use teacher-based information will not be as easy to design and implement as a new test, the pay-offs are immeasurable, as instruction focuses on the critical goals of literacy. If an initial worry about cost and logistics overrides such efforts, the more persistent concern remains -- will the critical goals of literacy be achieved by any but a select few?

References

Amarel, M., & Chittenden, E.A. (1982, June). *A Conceptual Study of Knowledge Use in Schools*. Final report to the National Institute of Education, ETS.

Anderson, R.C., Au, K.H., Borko, H., Guthrie, J., Hiebert, E., & Mason, J. *Teaching Reading Comprehension: Experience and Text*. Champaign, IL: Center for the Study of Reading.

Anderson, R.C., Hiebert, E.H., Scott, J.A., & Wilkinson, I.A. (1985). *Becoming a Nation of Readers*. Champaign, IL: Center for the Study of Reading.

Applebee, A.N., Langer, J.A., & Mullis, I.V.S. (1989). *Crossroads in American Education*. Princeton, NJ: Educational Testing Service.

Archbald, D.A., & Newman, F.M. (1988). *Beyond Standardized Testing* . Reston, VA: NASSP.

Assessment of Performance Unit (1987). *The Assessment of Reading*. Windsor, Berkeley, UK: NFER-Nelson Publishing Company Limited.

Berliner, D.C. (1987). *A Laboratory Science Component for Teacher Education Programs*. Tucson, AZ: University of Arizona.

Brewer, R. (1989, June). *State Assessments of Student Performance: Vermont*. Presentation at the 19th Annual Assessment Conference sponsored by Education Commission of the States, Colorado Department of Education. Boulder, CO.

Brown, R. (1989). Testing and Thoughtfulness. *Educational Leadership*, 46, 31-34.

Burstall, C. (1989, June). *"Integrated Assessment in the United Kingdom"*. Presentation at the 19th Annual Assessment Conference sponsored by Education Commission of the States, Colorado Department of Education. Boulder, CO.

Calfee, R.C. (1988). *Indicators of Literacy*. Santa Monica, CA: Center for Policy Research in Education, Rand Corporation.

Calfee, R.C., & Hiebert, E.H. (1988). "The Teacher's Role in Using Assessment to Improve Learning". In C.V. Bunderson (Ed.), *Assessment in the Service of Learning*. Princeton, NJ: Educational Testing Service.

Chapman, C. (1988, June). "Can Writing Assessment Improve Writing Instruction?" Illinois. Presentation at the 18th Annual Assessment Conference sponsored by the Education Commission of the States, Colorado Department of Education. Boulder, CO.

Clark, C.M., & Peterson, P.L. (1986). "Teachers' Thought Processes". In M.C. Wittrock (Ed.), *Handbook of Research on Teaching* (3rd Ed.). New York: Macmillan Publishing Co.

Dorr-Bremme, D.W., & Herman, J.L. (1986). *Assessing Student Achievement: a Profile of Classroom Practices (CSE Monograph #11)*. UCLA: Center for the Study of Evaluation.

Gil, D., Polin, R.M., Vinsonhaler, J.F., & Van Roekel, J. (1980). *The Impact of Training on Diagnostic Consistency (Technical Report No. 67)*. East Lansing, MI: The Institute for Research on Teaching.

Goodman, K.S. (1968). "The Psycholinguistic Nature of the Reading Process". In K.S. Goodman (Ed.) (1968). *The Psycholinguistic Nature of the Reading Process*, (pp. 15-26). Detroit: Wayne

State University Press.

Goodman, Y. (1985). "Kidwatching: Observing Children in the Classroom". In A. Jaggar & M.T. Smith-Burke (Eds.), *Observing the Language Learner*. Newark, DE: International Reading Association.

Gray, W.S. (1920). " The Value of Informal Tests of Reading Achievement". *Journal of Educational Research*, 103-11.

Hill, K.T. (1984). Debilitating Motivation and Testing: a Major Educational Problem--Possible Solutions and Policy". In R. Ames & C. Ames (Eds.), *Research on Motivation in Education: Student Motivation* (Vol. 1). New York: Academic Press, Inc.

Johnston, P.H. (1984). "Assessment in Reading". In P.D. Pearson (Ed.), *Handbook of Reading Research*. New York: Longman, Inc.

Marzano, R.J., Hagerty, P.J., Valencia, S.W., & DiStefano, P.P. (1987). Reading Diagnosis and Instruction: Theory into Practice. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Massachusetts Department of Education (1987). *Reading and Thinking: a New Framework for Comprehension*. Massachusetts Educational Assessment Program.

Meredith-Dabney, V. (1988, June). "Can Writing Assessment Improve Writing Instruction?", South Carolina. Presentation at the 18th Annual Assessment Conference sponsored by the Education Commission of the States, Colorado Department of Education. Boulder, CO.

Mosenthal, P., & Na, T.J. (1980). "Quality of Children's Recall under Two Classroom Testing Tasks: Towards a Socio-Psycholinguistic Model of Reading Comprehension". *Reading Research Quarterly*, 15, 504-528.

Popham, W.J., Cruse, K.L., Rankin, S.C., Sandifer, P.D., & Williams, P.L. (1985). "Measurement-Driven Instruction: It's on the Road". *Phi Delta Kappan*, 66, 628-634.

Reynolds, D. (1989, June). "State Assessments of Student Performance". Presentation at the 19th Annual Assessment Conference sponsored by Education Commission of the States, Colorado Department of Education. Boulder, CO.

Rindone, D. (1989, June). "State Assessments of Student Performance: Connecticut. Presentation at the 19th Annual Assessment Conference sponsored by Education Commission of the States, Colorado Department of Education. Boulder, CO.

Salmon-Cox, L. (1981). Teachers and Standardized Achievement Tests: What's Really Happening?" *Phi Delta Kappan*, 62, 631-634.

Scardamalia, M., & Bereiter, C. (1986). "Research on Written Composition". In M.C. Wittrock (Ed.), *Handbook of Research on Teaching* (3rd Ed.). New York: Macmillan Publishing Co.

Schafer, W.D., & R. Lissitz. (1987). "Measurement Training for School Personnel: Recommendations and Reality". *Journal of Teacher Education*, 38, 57-63.

Shepard, L. (1988, April). Should Instruction Be Measurement-Driven?" Opposing Views in Debate at the Annual Meeting of the American Educational Research Association, New Orleans.

Shulman, L.S. (1988). "A Union of Insufficiencies: Strategies for Teacher Assessment in a Period of Educational Reform". *Educational Leadership*, 45, 36-41.

Valencia, S., & Pearson, P.D. (1987). "Reading Assessment: Time for a Change". *The Reading Teacher*, 40, 726-732.

Valencia, S., McGinley, W., & Pearson, P.D. (in press). "Assessing Literacy in the Middle School". In G. Duffy (Ed.), *Reading in the Middle School* (2nd Ed.). Newark, DE: International Reading Association.

Vickers, D. (1988, June). "Can writing assessment improve writing instruction? South Carolina". Presentation at the 18th Annual Assessment Conference sponsored by the Education Commission of the States, Colorado Department of Education. Boulder, CO.

Wiggins, G. (1989). "Teaching to the (Authentic) Test". *Educational Leadership*, 46, 41-47.

Wixson, K.K., Peters, C.W., Weber, E.M., & Roeber, E.D. (1987). "New Directions in Statewide Reading Assessment". *The Reading Teacher*, 40, 749-754.

Wolfe, D. P. (1989). "Portfolio Assessment: Sampling Student Work". *Educational Leadership*, 46, 35-39.

Chapter 5

National Survey of the Use of Test Data for Educational Decision-Making

National Survey of the Use of Test Data for Educational Decision Making

Sheila W. Valencia, University of Washington[1]

The accountability movement of the 1970's, the many recent national reports (see Education Commission of the States, 1983 for a summary) and the focus of the effective schools research (Fisher, Berliner, Filby, Marliave, Cohen, Dishaw, & Moore, 1978) have set the stage for major educational reforms. In many instances, authors of these reports have relied on students' standardized test scores as measures of effectiveness or educational quality. Such a reliance has lead to an increased focus on testing: minimal competency testing, norm-referenced and criterion referenced testing. As a result, the use and potential influence of testing is greater now than at any time since World War I (Pipho, 1985). Evidence of the increasing use of tests is apparent from the 45 statewide competency testing programs now in place (Afflerbach, this volume). Add to this the thousands of locally regulated testing programs, the criterion-referenced tests accompanying every basal reading program, and the countless number of school and classroom tests, and the picture of a nation of schools, teachers and students engulfed by tests is complete.

Proponents of large-scale testing programs claim that a testing program can become a major force in improving classroom instruction (Haney, 1985; Popham & Rankin, 1981). They suggest that programs which use test results to drive instruction and instructional decision-making are exhibiting positive results (Popham, Cruse, Rankin, Sandifer, Williams, 1985). They find that when testing programs focus on significant competencies and student performance is tied directly to instructional consequences, tests drive curriculum in a most beneficial way. Not only do test motivate students and teachers, but they "remind" teachers of the focus of instruction and then provide important feedback on student progress.

In contrast, opponents argue that tests should follow rather than lead curriculum (Berlak,

---

[1] Conducted while the author was a Senior Research Scientist at the Center for the Study of Reading, University of Illinois at Urbana-Champaign, this paper was sponsored in part with funds from the Office of Educational Research and Improvement to the Reading Research and Education Center at the University of Illinois, Urbana-Champaign and from the Research Board of the University of Illinois, Urbana-Champaign. The opinions and conclusions expressed do not represent those of the funding agencies.

1985). They claim that overreliance on test scores leads to a narrowing of the curriculum, teaching to the format of the test rather than focusing on concepts and deep learning, and an emphasis on lower level, more easily tested skills (Linn, 1985; Madaus, 1985). They also point out that the results may be spurious; that we might develop a false sense of security from observing test gains that do not represent true growth in learning (Koretz, 1988, Valencia & Pearson, 1987). Madaus (1985) suggests that we are faced with tests which are so generic and curriculum insensitive that they are virtually useless. He finds that because of the commercial nature of the testing industry and because we have thus far resisted state or federally mandated curriculum, these tests have become so broad that they are unable to yield any useful information to guide instructional decisions. Opponents also remind us of the continual outcry from teachers concerning the disproportionate amount of time devoted to testing and the limited time available for instruction (Bridgman, 1988; Ordovensky, 1983; Ruddell & Kinzer, 1982).

Whether one supports or opposes the extensive use of test data for educational decision-making, many view the pervasiveness of testing as a fact of educational life: "Tests are likely to remain tools of policy implementation for the foreseeable future" (Madaus, 1985). Tests do shape (and derive from) educational policy and decision-making. In turn, they may shape curriculum at various levels of schooling (i.e. state, district, school, classroom). However, we have yet to develop a clear understanding of the nature of the influence of tests and test data on educational decision-making.

Relatively few studies have examined the impact of standardized test data on educational decision-making. Studies conducted in Pennsylvania, Ireland, and nationally suggested that neither teachers or administrators used standardized tests for classroom or curricular decision making (Burry, Catterall, Choppin & Dorr-Bremme, 1982; Kellaghan, Madaus & Airasian, 1980; Salmon-Cox, 1981; Sproull & Zubrow, 1981). More recent findings (Dorr-Bremme & Herman, 1986) indicate that teachers and administrators do use formal, standardized test results predominantly to report to others beyond the school level. They also rely, to a lesser degree, on this information for curricular decisions, planning, and placement although interviews suggest that this often involves a superficial and cursory examination of the results.

Alternatively, others (Brewer, Chambliss & Calfee, 1987) find that standardized tests do not affect most "mainstream classroom practices" but do exert a powerful influence on remedial and accelerated classes. At these extremes teachers are very concerned that students pass the minimum competency or advanced placement exams which represent expectations for those classes.

Still other studies show that in-class assessments such as teachers' tests and classroom observations are considered more valuable than standardized tests by teachers for instructional decisions; they provide teachers with the most immediate, instructionally relevant, and useful information (Brewer, Chambliss & Calfee, 1987; Dorr-Bremme, 1983; Dorr-Bremme & Herman, 1986; Gullickson, 1984; Haertel, Ferrara, Korpi & Prescott, 1984). Furthermore, teacher designed tests were found to align closely with instruction, but not necessarily with curriculum. That is, teachers made sure their tests measured what was taught but this was not always the same as the curriculum or course objectives (Haertel, in press).

The existing data base provides a beginning for understanding influence of assessment on educational decision-making and practice, but it predates the influence of most of the "Commission" reports. Since then, educators have called for research on the use of tests and the impact of tests on students, teachers and school districts (Rothman, 1988; Wallace, 1985; Madaus, 1981). They claim that without these data it is impossible to determine if measurement-driven instruction is a reality or myth. In essence, we are missing the link between classroom instruction and assessment. As noted by Ravitch, "There have always been lots of critics of tests, and lots of research on curriculum. But the two were looked at as separate issues. Now people have begun putting together discrete pieces of information, and asking whether or how tests drive curriculum."

An important additional factor must be considered--the perceptions of administrators and teachers. While accurate descriptions of the use of tests and test results is critical, they depict only a portion of the situation. We must come to understand what educators believe to be the use of test data. A comparison of actual use with perceived use may allow us to uncover the motivation behind the use of test results. Without considering these complementary perspectives we cannot understand the true impact of tests on instruction and curriculum. Findings from this study will provide the basis for a discussion from both perspectives.

Purpose

The study reported here focuses on testing in elementary and secondary schools in the United States. While several research questions targeted testing in general, the major questions in this study pertained to reading tests. Of the 45 states currently requiring state-wide assessment, all include a test of reading (see Afflerbach, this volume). Additionally, all standardized achievement batteries, which are used in some capacity in every state, include a reading section. If we are to investigate the impact of testing, reading seems to provide a logical domain: it is widely tested, an integral part of every curriculum, and a continual source of discussion nationwide.

The research focused on three goals: The first goal was to obtain an accurate portrait of the scope and nature of testing in general, and reading testing in particular, in U.S. schools. A second goal was to determine how reading tests and test data influence the actions of teachers and administrators. The third goal was to compare the actual and perceived use, frequency, and impact of reading testing on the decisions of teachers and administrators in an effort understand the forces behind those decisions.

Method Instrumentation

In June, 1986 a written mail questionnaire was developed for administrators and teachers using a three-stage process. First, after a review of the literature and existing questionnaires, a conceptual framework was identified. The framework consisted of five general questions:

- What is the nature of the general achievement testing program in U.S. schools?
- What is the nature of reading evaluation in U.S. schools?
- How do reading tests and test results influence classroom and administrative decisions and practices?
- How are reading data used by teachers and administrators?
- How do teachers' and administrators' perceptions of testing compare with their actual reports and their recommended practices?

Open-ended items were constructed under each of the major questions and then administered to a group of approximately 25 teachers and administrators. An analysis of their responses suggested possible options for inclusion in the constructed response format to be administered to the larger sample. Next, a paper/pencil survey was developed and piloted with approximately 40 teachers in a large metropolitan school district. These teachers responded to each question, suggested additional or alternative options, and critiqued the format and directions. At the same time, the survey was reviewed by four testing and evaluation experts. Based on the feedback, the survey was revised and two different forms were constructed--one for administrators and one for teachers. The framework, format, and question stems were identical for both forms. However, several of the options were altered slightly to address the variability in job responsibilities.

Subjects

A complete listing of all school districts in the United States was obtained from a data tape provided by the National Center for Educational Statistics. A stratified systematic sample of 10% of the 14,535 school districts was selected for inclusion in the study with some oversampling designated in the smaller cells. In all, 1,475 school districts were selected representing the strata of enrollment size (<1,000, 1,000-24,999, >25,000), location (urban, suburban, rural), and type (K-8, K-12, 9-12).

Procedure

School districts represented the first level of sampling. In October, 1987 superintendents were requested to complete the administrator questionnaire and to return a list of names and addresses of schools within their district. A follow-up superintendent mailing was sent three weeks later.

From the list of schools, one elementary and one secondary school were randomly selected for inclusion in the study. Principals were sent the administrator questionnaire and five teacher questionnaires to be returned directly to the research office. Principals were asked to randomly distribute the questionnaires to teachers in their buildings. English/Language arts teachers participated at the secondary level, and a combination of primary and intermediate grade teachers participated at the elementary level. Follow-up mailings were sent at weeks 3 and 5. Response rates averaged 37% overall and 45% across district level sampling strata.

Analyses

The distribution of the responses by location and enrollment corresponded very closely to the distribution of the sample, and thus to the population. Therefore, all responses were used in the analyses. Data were analyzed using the Loglinear-maximum likelihood (LLM) approach. A reduced main effects model was tested and found to be an adequate fit to the data ($p < .25$). Hypotheses tested concerned main effects for level (elementary vs. secondary), job (administrator vs. teacher), and enrollment size (small--<1,000 vs. medium--1,000-24,999 vs. large-->25,000).

Results

Responses were received from a total of 1890 administrators and teachers representing 543 school districts in 50 states. All appeared to have been experienced educators, reporting an average of 21.7 years experience for administrators and 15.4 years for the teachers.

General Achievement Testing

Type and frequency of achievement testing. The first series of questions pertained to the nature of general achievement testing in U.S. schools. As we might anticipate, more than 94% of all respondents reported that standardized norm-referenced tests were administered and an average of approximately 65% reported that state-mandated tests were also administered. Less than 20% administered district-developed tests; these types of tests are most often associated with objective-based management systems and resemble criterion-referenced tests designed to align with the specific outcomes/objectives identified by a school district.

While there is little difference between the amount of standardized testing reported by elementary and secondary teachers, there is a significant difference between their reports of state-mandated testing; significantly more secondary educators report state-mandated testing than elementary educators. This discrepancy most likely can be attributed to two factors: 1)the presence of many statewide minimal competency types of exams required for graduation, and 2)the likelihood that the primary grades (K-3) are not as frequently involved in statewide testing (Teale, this volume). The similarity in administrators' reports at the elementary and secondary levels reaffirms that the difference is probably a reflection of the difference between teachers' isolated classroom perspective and a more global district view.

On average, both teachers and administrators estimated that 6-7 hours were spend administering the norm-reference tests and 2-3 hours administering the state-mandated tests. As a conservative estimate, norm-referenced testing accounts for approximately .7% of the annual academic instructional time; a relatively small amount of time for actual test administration.

In general, although there are some subtle differences across groups, the main findings indicate no significant differences for job (teacher vs. principal), level (elementary vs. secondary),or enrollment (small vs. medium vs. large schools) concerning opinions about the amount and use of general achievement test data in U.S. schools (See Table 1). The majority of respondents indicate that "the right amount" of testing is taking place at the state, district and school levels but at the national level only 35% believe it is "the right amount" and an additional 19% believing there is "too much" national testing. Almost 30% of the sample had no opinion, or felt that they did not have sufficient information, to respond to the question about nationwide testing. It is interesting that as the question moves from a wider scope (the entire U.S.) to a more specific perspective (individual teachers' classrooms), there is an increase in the percentage of teachers who believe that "the right amount" of testing is taking place. It appears that as testing becomes closer to classroom life and thus, to teachers' first-hand experiences and to their control, teachers seem to be more certain that the correct amount of testing is taking place.

Use of test data. Two questions were asked about the actual and optimal use of test data. The first sought to obtain information about existing practices and uses of test information, and the second was targeted at determining if those practices were perceived to be appropriate. Several interesting trends emerge from these data. First, there are significant differences between responses of principals and teachers for almost all the uses actually engaged in (curriculum revision, instructional decisions, evaluation/ranking). Specifically, more teachers report that data are used for evaluation/ranking of states, districts, schools, and teaching staff than do administrators. In contrast, significantly more administrators than teachers report that data are used for decision-making regarding students, instruction and curriculum. Although 43-78% of all respondents indicate that data are used for various types of classroom decision-making, it is interesting that these kinds of uses are reported by more administrators than teachers. It would appear that there are misperceptions on both sides of the question: administrators believe that testing data are more widely used and useful in classroom decision-making than do teachers, yet teachers believe that testing data are more widely used for administrative ranking and evaluation purposes than do administrators.

Table 1

What is your opinion of the amount of testing that currently is being carried on in:

	Elementary Principals	Elementary Teachers	Secondary Principals	Secondary Teachers
THE ENTIRE UNITED STATES				
Too much	24.24	20.96	13.04	17.13
The right amount	34.55	31.99	37.68	34.26
Too little	4.85	8.82	13.04	12.96
YOUR STATE				
Too much	31.79	22.67	22.73	15.30
The right amount	47.98	49.18	56.20	56.85
Too little	6.94	7.68	10.95	11.19
YOUR SCHOOL DISTRICT				
Too much	22.73	20.00	16.67	12.70
The right amount	63.64	64.82	71.01	64.63
Too little	11.36	9.29	11.59	16.33
YOUR SCHOOL				
Too much	22.41	18.92	18.38	11.82
The right amount	65.52	68.65	67.65	66.82
Too little	10.34	9.55	13.24	17.05
YOUR CLASSROOM				
Too much	NA	15.50	NA	4.85
The right amount	NA	75.68	NA	77.14
Too little	NA	6.13	NA	9.24

Thus, both teachers and administrators do not appear to rely on the test data as much as others believe they do for purposes most directly related to their job responsibilities

The second trend is apparent from a comparison of the reports of actual use of test data and the recommended use of test data. A fairly small percent (12%-29%) of teachers and administrators believe that test data should be used for evaluating and ranking of states, school districts, schools, or teachers. This is a major change from the actual use where 43- 65% reported that data were used to evaluate states, districts and schools. (Very few administrators or teachers reported using data to evaluate teachers). The discrepancy seems to indicate some dissatisfaction with the competitive comparisons reminiscent of Secretary Bell's wallchart and forecasted by the impending state-by-state NAEP comparisons of the 1990s. Additionally, there is an overall perception held by an increasing number of both administrators and teachers that we should increase the use of these data for instructional decision-making. Although a greater percentage of both groups selected these instructional uses (e.g. decisions about students, teaching, curriculum), the significant difference between administrators and teachers regarding the actual use of test data prevails in questions about recommended use; more administrators favor using test data for classroom decision-making than do teachers.

In summary, the broad picture we get of the general achievement testing program in U.S. schools is one where standardized testing is common practice. There is widespread use of test data for comparative evaluation of states, districts, and schools. However, many more teachers and administrators express a desire to use the data for instructional decisions.

There are also significant differences in the use of achievement data by teachers and administrators with each group reporting greater use for others' purposes than forthemselves. These data highlight the concern of many educators regarding obtaining useful information from tests. It raises several recurring the issues-- can or should a single instrument be able to serve many different purposes, and is it appropriate to use standardized tests to make decisions about instruction and instructional programs? (Cole, 1988; Johnston, 1987; Cross & Paris, 1987; Valencia & Pearson, 1987) These issues are presented in more detail in the discussion section of this paper.

Reading Tests

The majority of the survey focused specifically on reading tests. The questions were designed to help us understand how these tests and test data influence curricular and instructional decisions about reading instruction. First, respondents were asked to report how often they engaged in seven different types of reading evaluations. Table 2 lists the types of reading evaluations most commonly found in schools from those most likely to be externally imposed to those more internally imposed and most tightly aligned with instruction.

Once again, commercially published norm referenced tests and state-mandated tests are reported as administered once or twice a year by a majority of elementary and secondary principals and teachers. These externally imposed evaluations seem to be a consistent but relatively small part of the reading evaluation picture when compared with the frequency of other measures such as basal reading tests and classroom assessments.

In fact, the amount of time teachers spend reviewing or specially preparing students for standardized and state-mandated tests is actually very limited (approximately 4-7 hours per year--See Table 3). Taken together with the information above about time spent administering standardized tests, the time commitment still figures to be less than 2% of the annual academic instructional time.

The predominant mode of reading evaluation appears to include measures that are more classroom and curriculum-based (e.g. basal reading series tests, teacher made tests, written assignments, classroom observations). Basal reading tests are curriculum embedded tests which accompany all commercially published reading textbook series. At a minimum, they are used as benchmarks 3-5 times during the

Table 2

How often are the following types of reading evaluations admnistered to students for whom you are responsible?

Times per year (% of teachers)

Type of Test	Elementary Principals	Elementary Teachers	Secondary Principals	Secondary Teachers
Commericial	1-2 (83)	1-2 (79)	1-2 (79)	1-2 (65)
State-mandated	1-2 (78)	1-2 (50) 0 (47)	1-2 (67)	1-2 (61) 0 (36)
District-constructed	0 (59)	0 (82)	0 (69)	0 (80)
Basal reading series	9+(29) 3-8 (58)	9+(22) 3-8 (59)	1-2 (20) 0 (58)	1-2 (10) 0 (67)
Teacher-made	9+(42) 3-8 (35)	9+(37) 3-8 (27)	9+(42) 3-8 (23)	9+(42) 3-8 (22)
Written assignments	9+(66)	9+(71)	9+(58)	9+(75)
Observations	9+(48) 3-8 (30)	9+(77) 3-8 (11)	9+(40) 3-8 (27)	9+(62) 3-8 (19)

Table 3

Within a school year approximately how much time do you think YOU SPEND (or TYPICAL TEACHERS SPEND) reviewing or specially preparing students to take the following types of reading tests?

Type of test	Elementary Principals	Elementary Teachers	Secondary Principals	Secondary Teachers
Commercially published	6.91	6.74	3.30	4.17
State-mandated	7.32	6.38	5.51	4.51
District-constructed	2.04	1.78	2.15	1.19
Basal reading series	13.77	26.83	5.97	6.03
Teacher-made	13.99	18.06	15.08	16.90

school year to measure progress through the program. At a maximum, some series provide pre-tests and post-tests for every skill taught in the series. It is estimated that basal reading series account for 75-90% of the reading instruction in elementary classrooms nationwide (Anderson, Hiebert, Scott & Wilkinson, 1985), so it is not surprising that many teachers report using the tests which accompany them. However, the overwhelming percentage of teachers who use basal tests and the frequency with which they are used is most surprising.

More than 80% of the elementary teachers use basal tests 3 or more times a year and 22% of this group use them more than 9 times per year. In this category elementary principal's perceptions are fairly close to reality.

In contrast, the data for secondary schools is less revealing. Because basal textbooks are predominantly an elementary school phenomenon, we would expect, and actually do find, that few teachers reported basal test use at the secondary school.

Teacher-made tests, written assignments and observations are all widely used for evaluation at both the elementary and secondary levels. However, once again, as indicated in the results of questions about general achievement testing, there is a discrepancy between teachers' reports and principals' perceptions. With the exception of teacher-made tests at the secondary level, principals consistently significantly underestimate the frequency and time spent by teachers preparing students to take internally controlled classroom evaluations.

There may be several explanations for these discrepancies. One explanation may be that principals may not define evaluation in the same way as teachers; that is, they may not classify many of the more interactive, classroom tests and activities as evaluation. They may define evaluation more narrowly, as more formal tests. An alternative explanation may be that definitions are, in fact, similar yet because administrators spend little time in the classroom, they may not be aware of how often teachers use these less formal modes to assess students. In either case, it is clear that teachers report engaging in much more classroom based evaluation than administrators believe they do.

The influence of tests on decision-making. The considerable presence of testing in our schools is demonstrated by the data from this study. But an equally important corollary pertains to how the results of these tests influence the actions of teachers and administrators and how they are used to shape decisions.

The influence of tests on educators can be conceptualized in terms of actions and thoughts. In the former case, the presence, or threat, of tests might encourage teachers or administrators to adapt their actions to assure optimal student performance. For example, teachers who are aware that a standardized test will be administered in February and will include questions about a particular skill or content may adapt instruction to be sure that topic is covered before the testing date.

Sometimes, the influence may not be as direct; simply the concern of an impending test may motivate a week of review or rehearsal of test taking strategies. Additionally, the results of tests might influence priorities or resource allocations for material or programs aimed at preventing low scores or shoring up deficiencies.

In the latter case, the influence of tests on educators' thoughts, tests might influence expectations. Teachers and administrators may use tests as a bellwether of the academic ability of students. Although such a use may help

Table 4

How do standardized/norm-referenced reading tests and their results influence what you do?

	Elementary Principals	Elementary Teachers	Secondary Principals	Secondary Teachers
I alter what I teach to be sure I cover what is tested.	NA	37.26	NA	26.55
I spend time with my students practicing/ reviewing for tests	NA	48.86	NA	39.38
Help me to know how much to expect from individual students/ teachers	65.56	54.37	64.08	52.10
I teach my students/ teachers test-taking strategies	36.11	56.99	30.28	62.03
Help me prioritize and set goals for the year	72.78	59.44	60.56	52.32
I allocate more or less of my resources depending on the scores	27.22	27.80	26.76	22.08
I suggest/implement curricular changes	73.33	43.71	75.35	48.34
I provide inservice, teacher support, and supervision	72.22	NA	62.68	NA
Tests and test results do not influence me	3.89	10.14	2.11	10.60

suggest broadly defined needs and accomplishments, it is always accompanied by the potential danger of becoming a self-fulfilling prophesy. Thus the influence of tests may foster both positive and negative actions and thoughts on curriculum and instruction.

The data on the influence of tests seems to run contrary to popular belief in one respect; only a small percentage of teachers report that they alter instruction to match the test (See Table 4). However, an increasing number of elementary and secondary teachers (39%-62%) report spending time reviewing for tests and actually teaching test-taking strategies. So although the content of tests doesn't appear to influence classroom instruction, testing does influence instructional time in by introducing special review and preparation for the testing experience. In every instance of these instructional influences, elementary teachers report being more influenced than secondary teachers.

There may be several factors that account for the appearance of a limited impact of testing on instruction. First, the avenue of impact is often most keenly felt at the district or school level. That is, school districts or building administrators may respond to test content by reviewing and revising district goals or objectives, in other words, rethinking curriculum. These objectives, in turn, are passed along to teachers as expectations. Thus, teachers may not see the need to alter the content of their instruction to match a test because, in essence, the curriculum has already been changed and passed along to them.

A second explanation for the minimal influence on classroom instruction may be found in the content of most reading tests. Most standardized group reading tests are fairly generic--they include a vocabulary and comprehension section. Only a very few lower grade tests still include subtests of discrete reading skills such as decoding strategies and reference skills. Therefore, most reading programs and instructional procedures will foster proficiency in the two major areas of vocabulary and comprehension. It may be that reading tests are more global than other content area tests and therefore reading instruction and curriculum may be less susceptible to test-driven influence than, for example, mathematics, science or social studies.

A third possibility is that teachers may not have access to the actual tests or information on the test coverage before the test is administered. In this case, the lack of influence would be attributed more to lack of information than lack of desire to alter coverage.

A final possibility is that teachers simply do not let tests influence their instruction. However, given the fairly wide reports of review and preparation for tests, it would seem that teachers are concerned and do take steps to adapt to the requirements of testing. Therefore, the first two explanations, the paths of indirect influence, seem most plausible.

Table 4 also reveals information about administrators' actions; a great percentage of them report that test results influence their decisions about curriculum and help them determine needed support and staff development for teachers. Additionally, significantly more administrators than teachers use test data to set priorities and goals for the academic year. However, administrators do not appear use test results to differentially allocate resources. These findings support the hypothesis above that administrators are the agents who shape curriculum based on test results and that teachers seem to be the recipients of those adaptations.

In terms of influence on educators thoughts, a majority of teachers and administrators report that test results clearly influence their expectations for students and teachers. However, there are significant differences between teachers and administrators. While more than 50% of the teachers report using test rests to guide expectations, a significantly greater percentage of administrators use results for that purpose.

The influence of standardized tests depicted from these data supports the notion that test results influence administrators who, in turn, shape curriculum. In contrast, the majority of teachers do not report much substantive change in classroom content coverage but do devote time to preparing students for the test taking experience.

The use of reading test data. The question of use of test data takes on two dimensions--that of actual use and of perceived use. The information on actual reported use is presented in Table 5.

Three obvious trends emerge from these data. First, the results of standardized reading tests are used by a substantial percent of educators. Apparently, teachers appear to consider test scores selectively, preferring the indicators of students' best performance. This is consistent with the emphasis on classroom-based assessment report by most teachers.
Although a small percentage of teachers use the information for comparative purposes and for grouping for reading instruction, a large percentage use it for making decisions about individual students (e.g. diagnosis and tracking) and, at the elementary level, for reporting to parents. This apparent discrepancy between different classroom uses (grouping for instruction vs. diagnosis and tracking) can probably be attributed to the wide use of basal reading series tests indicated in Table 2. These tests are more likely than standardized test to be

Second, significantly more elementary teachers and principals use this information than do their secondary counterparts. Third, for uses which apply to both teachers and administrators, significantly more administrators rely on these data than do teachers.

A substantial percent of teachers report using standardized test results as supporting or confirming information--as a supplement to other classroom evaluations. This is consistent with other studies (Madaus, 1985; Salmon-Cox, 1981) that found that teachers used test data to confirm their judgment and to guide their decisions. Interestingly, Salmon-Cox (1981) found that when test scores were lower than class performance, teachers tended to disregard test scores; when test scores were higher, teachers paid more attention to them.

used for placement and grouping since they are specific to an adopted reading program. The prevalent use of data to for diagnosis of individual students is consistent with Ruddell's (1985) findings that teachers want tests that help them diagnose individual students' needs. The important question however, that is discussed below, is whether group standardized tests can reliably provide that information. And, interestingly, when asked what might encourage teachers to use standardized reading results more than they do at present, approximately 42% suggested that these tests don't measure what they should or what is actually taught, nor do they offer the teacher any new information. On one hand, teachers want more diagnostic information from tests but, on the other hand, they doubt the validity of these very same measures.

There is a consistent, and sometimes significant, difference between elementary and secondary teachers for every test result use listed in Table 5. Significantly more elementary teachers rely on tests more for every use listed reporting to parents, confirming progress, grouping, and special placements. This may reflect both a true difference in use as well as the difference in the organizational structure of the levels. Secondary teachers see 5-6 times the number of students most elementary teachers see and often administer and receive results for students in their homeroom rather than those in their instructional classes. Thus, use of reading test data becomes less directly a part of secondary teachers instructional responsibilities and therefore less clearly useful for them.

The actual use of data for administrators depicts even wider use than it does for teachers. Again, the vast majority of administrators use test data for estimating students' ability, tracking students in special academic programs, and reporting to parents. Parallel to the teacher data, many fewer administrators use test results to make comparisons among students, schools or districts. Additionally, test results are used by most administrators for administrative tasks such as goal setting, program evaluation and teacher conferencing. Very few principals report using the information as a determinant for teacher reassignment or reward.

Overall, as noted above, for all types of uses reported by 50% or more of the respondents, more elementary teachers and principals report using standardized reading test results than do secondary educators.

A similar set of use questions was asked to explore perceived uses. That is, teachers were asked about the purposes for which the "typical administrator" used test results, and administrators were asked about the "typical teacher." A comparison of actual teacher uses with perceptions of teacher uses discloses some interesting contrasts. In general, more administrators believe that teachers use standardized reading test data than actually reported by teachers. Administrators seem to believe that test data are more useful for classroom decisions, reaffirming evaluations and communicating with parents than do teachers. The one striking reversal is in the area of making comparisons; more teachers, although relatively few (22%-35%) actually make use test data for comparative purposes than principals believe (12%- 13%).

The analogous comparison for administrators is also interesting. Once again, the perceptions indicate greater use than the actual reports. More teachers believe that administrators use test data for comparisons and goal setting than they actually do; however fewer report the occurrence of discussions between administrators and teachers concerning improving test scores, a more instructionally based use, than administrators actually report.

The patterns of discrepancies between perceptions and reality seem to suggest that educators believe that test information is more

Table 5

Indicate whether you use standardized/norm-referenced reading test results for each purpose:

	Elementary Principals	Elementary Teachers	Secondary Principals	Secondary Teachers
I use the information to group for reading instruction	NA	47.63	NA	21.68
I use the information to confirm my evaluation of student progress	NA	70.47	NA	49.56
I use the information to diagnose individual student difficulties	NA	69.07	NA	63.72
I report results to parents	82.22	67.66	73.05	24.28
I get an idea of students' abilities	90.56	77.10	86.52	72.19
I use the information for referrals/tracking students	82.78	65.21	83.69	50.99
I compare the results of my students with other students/schools/districts	43.33	35.14	36.88	21.63
I talk with teachers about their class scores and how to improve them	63.48	NA	42.14	NA
I set goals for the school based on test scores	55.06	NA	48.57	NA
I set goals for individual teachers based on test scores	20.22	NA	16.43	NA
I use the information as one of several bases for teacher reassignment	8.99	NA	11.43	NA
I use test scores to determine the effectiveness of the school/district reading program	79.21	NA	58.57	NA
I recommend teachers for salary increments, merit awards, and promotion based, in part, on test scores	1.12	NA	0.00	NA
I do not use them at all	0.56	4.90	2.13	10.38

useful to fulfill others' needs than the "others" believe. More specifically, teachers believe administrators use information more for programmatic evaluation than they actually do but also believe that they use it less for instructional improvement than administrators do. Likewise, administrators believe teachers use the data for classroom decisions more than they actually do but believe teachers use it less for comparisons outside the classroom than teachers report. It appears that each group is using test data more for others' purposes than for their own direct responsibilities and that both administrators and teachers believe test results are more useful to others than they actually are.

This situation might be influenced by teachers' and principals' unfamiliarity with one another's evaluation tools. Simply put, each may not be aware of the assortment of tools used by the other and thus place more emphasis on the commonly known tool, the standardized test, than does the actual user. Even with these discrepancies though, the fact remains that most administrators and teachers are relying on standardized tests to make important instructional and programmatic decisions.

Purposes of reading tests. Another perspective on uses was obtained by asking respondents to indicate their understanding of the broad purposes for which reading test results ARE appropriately used and their opinions about the purposes for which results SHOULD BE appropriately used (See Tables 6 & 7).

The data indicate very low actual and desired utility of results for administrative control (e.g. funding of educational programs and materials, teacher monitoring, etc.). However, the desire for useful instructional information is once again evident in these responses. Most elementary and secondary teachers and principals report that results are appropriately used for curriculum development and instructional planning and even a greater percentage believe that results should be appropriately used more than at present. While more secondary teachers would like to see reading tests used more for promotion/retention, fewer elementary teachers believe they should be used for this purpose.

A final perspective on the uses of test data was provided when respondents indicated the types of tests which they believe are useful for different purposes. Specifically, the questions examined whether administrators and teachers believed that tests had differential utility for decisions about special reading programs, grouping/tracking, group instruction, and individual instruction.

The main finding from these questions is that educators have very clear ideas about the utility of various tests. There is overwhelming agreement from most principals and teachers that standardized norm-referenced reading tests are useful for referral to special programs (approximately 87%) and for grouping or tracking students within a school (approximately 61%). Once again, however, administrators seem to rely on reading test results significantly more than teachers. In addition, a majority of elementary educators see the basal reading reading tests as useful for referrals and tracking as well. This is an interesting combination of relying on national comparative data as well as specific classroom data for making decisions to special programs.

The magnitude of dependency on standardized reading test results is probably a reflection of the requirements imposed by many state and federally funded special programs and the availability of standardized test results stemming from their widespread use in schools.

Table 6

What do you think are the appropriate uses of results from reading tests?

	Elementary Principals	Elementary Teachers	Secondary Principals	Secondary Teachers
Make decisions about funding	16.67	21.68	21.13	18.76
Salary increments (merit awards for teachers)	1.11	3.32	0.70	2.21
Teacher evaluation	6.67	12.59	7.04	9.05
Teacher dismissal	1.67	2.80	2.82	1.99
Curriculum development	75.56	59.44	78.17	59.16
Plan classroom instruction	69.44	60.81	75.35	57.40
Promotion/retention of students	57.22	53.15	44.37	35.10

Table 7

What do you think *should be* the appropriate uses of results from reading tests?

	Elementary Principals	Elementary Teachers	Secondary Principals	Secondary Teachers
Make decisions about funding	16.11	16.43	22.54	20.53
Salary increments (merit awards for teachers)	4.44	3.67	2.82	3.31
Teacher evaluation	9.44	4.55	7.75	3.75
Teacher dismissal	1.11	1.75	4.93	1.32
Curriculum development	86.11	80.94	85.21	76.60
Plan classroom instruction	82.22	72.03	82.39	73.51
Promotion/retention of students	54.44	47.03	44.37	50.11

BEST COPY AVAILABLE

The patterns of responses for the two questions about group instructional decisions and individual student instructional decisions are almost identical to one another. There are three main findings: a majority of principals report that standardized tests are useful for classroom instructional decisions but many fewer teachers concur; elementary educators find basal reading series tests useful for instructional decisions; and a majority (approximately 63%) of teachers and administrators use teacher- constructed reading tests for instructional decisions although more elementary personnel use them than do secondary. It is interesting that there is little differentiation between group and individual decisions, even with respect to the use of standardized measures, which most often do not yield diagnostic data.

The striking lack of utility of state and district constructed tests for referral, tracking, or instructional decisions suggests that these measures have very specific purposes and audiences which are neither programmatic or instructional. These findings, combined with the from Table 2 would seem to assert that data from these sources are probably targeted at state and district administrators who are more removed from school and classroom decisions. School level people clearly do not rely on these types of tests, they simply administer them as required.

Other reading assessment strategies. Finally, there is very strong evidence that most teachers use informal classroom observations, performance assessments, and review of classroom assignments, in addition to tests, to assess their students. More than 85% of all teachers and administrators report using observations and classroom assignments and more than 64% report using performance assessments to evaluate students. Overall, more elementary teachers tend to use classroom observations and projects for evaluation than do secondary teachers. Additionally, although not significant, there appears to be a trend for elementary administrators to underestimate teachers' use of these informal measures and for secondary administrators to overestimate their use.

Discussion

The findings from this study suggest that standardized and state-mandated tests in general, and specifically in reading, are a constant component of the school assessment process. In a majority of schools, they influence decisions about educational goals, curriculum, expectations, and time spent preparing for tests, but do not seem to exert such a powerful influence on classroom instruction. Reading test results are frequently used to communicate with people outside the school (e.g. parents, funding agencies, special programs), to evaluate reading programs, to support other assessment information, and to make diagnostic decisions regarding students. There is widespread use of reading test data for curricular and instructional planning and an overwhelming belief that this information should be used even more than it is.

On the other hand, the most widely used assessments are those that are less formal, classroom controlled, and directly aligned with instruction (e.g. observations, classroom projects, basal reading tests). These take precedence in planning for instruction over both more formal classroom assessments (teachers' tests) and standardized tests. The vast majority of teachers rely heavily on these types of assessments to guide instruction. However, administrators are consistently unaware of the magnitude of teachers' reliance on classroom-based evaluations.

There are consistent, significant discrepancies between reports by administrators and teachers. In some cases, this represents a difference in their roles and their need for test results. Overall, administrators depend on standardized test results more than teachers. This seems logical, given that administrators often must report results for large groups of students in a consistent and efficient manner. However, both administrators and teachers see the information as more useful for others' purposes than for their own.

The findings from this study help illustrate several points. First, it appears that although both administrators and teachers use standardized tests, use of tests seems to focussed on purposes for which each group is least familiar. In other words, this may be an issue of security and trust. Principles are most familiar with (and responsible for) evaluation of curriculum and programs, yet a majority report that they use standardized test results, and expect that teachers use standardized test results for instructional decisions. Teachers do just the opposite. They report greater use of test scores for programmatic evaluation and less for instruction while maintaining a belief that administrators use the information for administrative decisions. Each group, administrators and teachers, seems to have more confidence in their own abilities than in others', perhaps reflecting an understanding the complexity of their evaluation responsibilities and relying on their own expertise in using other assessment strategies.

A second interesting trend suggests that the avenue of impact and influence of standardized reading tests may not be directly linked to instruction. There may be an indirect influence on instruction via district or published curriculum so that teachers are unaware of the true impact of standardized tests on their instruction. Alternatively, standardized tests may not have a substantial influence on instructional practices or content, but rather influence instructional time by making teachers feel that they must devote a portion of it to test-taking preparation. In some ways, this latter alternative may represent a more dangerous scenario than the former. Spending classroom time coaching students on test-taking strategies doesn't impact instruction or curriculum in a positive way, it simply prepares students for tests at the expense of meaningful learning (Madaus, 1985). Although the data from this study do not permit the disambiguation of the possible paths of influence, the findings do strongly suggest that standardized tests most assuredly impact the school curriculum and goal setting.

Many of the data suggest a third conclusion. It appears that teachers and administrators are trying to optimize the use of standardized test results by using the information for programmatic and diagnostic purposes. This is a logical desire for many who feel that if they are engaging in assessment exercises, they should be able to make use of the results (Cole, 1988; Dorre-Bremme & Herman 1986). However, the use of standardized test data for instructional and diagnostic purposes may pose some critical problems and potentially lead to misdirected conclusions and dangerous curricular and instructional decisions.

One of the assumptions behind the use of test data is that the users understand the value and limitations of the data; that they have a firm understanding of the concepts taught and assessed as well as a general sophistication in measurement concepts. However, it is important to note the misconceptions and inadequate training provided to educators in the area of testing and measurement and the resulting impact on the interpretation of test data (see Yeh, 1980 for a discussion). For example, a study conducted by Ruddell and Kinzer (Ruddell, 1985), revealed that only 11% of the teachers and 17% of the principals understood the concepts of scaled scores and standard error of measurement. Most educators in that study emphasized the need for assistance in the interpretation of information such as raw scores, percentiles, and scaled scores. Concerns such as these prompted the Delegate Assembly of the International Reading Association (April, 1981) to pass a resolution enumerating the misuses of grade-equivalent scores and advocating the elimination of the use of grade equivalents by educators and test publishers. Data from this study indicate that teachers and administrators are using results from standardized tests to make important decisions, yet we have reason to believe that these educators may not have the background or knowledge to make valid use of the information.

A related issue in understanding the limitations of test results is the use of standardized tests for diagnostic purposes. A standardized test may include a sufficient number of items for obtaining a "reliable global score" but may only contain 3 or 4 items on a particular objective. Any diagnostic decisions, either for the individual student or the curriculum, based on such few items are likely to be invalid (Linn, 1986; Wang, 1988).

Similarly, due to time constraints, tests must be selective in the objectives or concepts covered (Airaisian & Madaus, 1983; Rudman, 1987). Therefore it is very likely that many of the objectives stressed in the curriculum will not be tested (Aronson & Farr, 1988). This may make the interpretation of results for the purposes of specific district or school program evaluation very tentative. There are, in the current structure of standardized tests, three competing goals of testing: the normative assessment of large groups, the evaluation of specific programs, and the diagnosis of individuals (Airaisian & Madaus, 1983; Cole, 1988; Dorre-Bremme & Herman, 1986; Linn, 1986). The caveats above demonstrate the apparent tension between creating a standardized test that can be used by a broad sector of the country and one that is sensitive to individual student diagnosis and local curriculum. The problem seems to be that consumers believe that the same tests can serve all functions. It is from this belief and the desire to make use of available information that the potential for misuse emerges. Caution must be exercised by all who use standardized test results.

Finally, all of the issues above must be considered in the context of widespread concern of many reading educators about the validity of existing reading tests and the mismatch between reading assessment and instruction. Advances in reading research are being implemented in many classrooms and curriculum guides, but they have not been integrated into assessment instruments--most especially standardized tests (Afflerbach, this volume; Farr & Carey, 1986; Johnston, 1988; Valencia & Pearson, 1987). This would suggest that even if educators heeded all the cautions and used test results selectively, the validity of results would be suspect.

The juxtaposition of findings from this study with the cautions of measurement and reading educators present a dichotomy. It is clear that standardized testing is rampant in the United States and that it is likely to maintain, if not increase its impact on schools (Linn, 1986, Pipho, 1985; Wang, 1988). If standardized testing occupies such a strong position in schools and classrooms, then one would hope that the information gleaned from these instruments would be meaningful and useful. On the other hand, if we are concerned about the appropriateness of many reading measures, and cautioned about the use of these data for programmatic, instructional, and individual decisions, then it imperative that we suggest and communicate clear guidelines for the use of test results. If education is to benefit from the current emphasis on testing, there is much work to be done.

References

Afflerbach, (this volume) Statewide Reading Assessment: A Survey of the States.

Airasian, P.W. & Madaus, G.F. (1983). Linking testing and instruction: Policy issues. *Journal of Educational Measurement*, *20*, 103-117.

Anderson, R. C., Hiebert, E. H., Scott, J. A., & Wilkinson, I. A. G. (1985). *Becoming a Nation of Readers: The Report of the Commission on Reading*. Washington, DC: National Institute of Education.

Aronson, E. & Farr, R. (1988). Issues in assessment. *Journal of Reading*, *32*, 174-177.

Berlak, H. (1985). Testing in a democracy. *Educational Leadership*, *43*, 16-17.

Brewer, A., Chambliss, M., & Calfee, R. (1987). The Impact of assessment on instruction. Paper presented at the annual meeting of the American Educational Research Association.

Bridgeman, A. (1985). Standardized test deficient, English teachers assert. *Education Week*, Dec. 4, 1985, p.4.

Burry, J., Catterall, J., Choppin, B., & Dorr-Bremme, D. (1982). *Testing in the nation's schools and districts: How much? What kinds? To what ends? At what costs?* CSE Report No. 194. Los Angeles: UCLA Center for the Study of Evaluation.

Cole, N. S. (1987). A Realist's appraisal of the prospects for unifying instruction and assessment. In *Assessment in the service of learning: Proceedings of the 1987 ETS Invitational Conference*. Princeton, NJ: Educational Testing Service.

Cross, D. R., & Paris, S. G. (1987). Assessment of reading comprehension: Matching test purposes and test properties. *Educational Psychologist, 22(3 & 4)*, 313-332.

Dorr-Bremme, D. W., & Herman, J. L. (1986). *Assessing student achievement: A profile of classroom practices*. Monograph No. 11, CSE Monograph Series in Evaluation. Los Angeles: UCLA Center for the Study of Evaluation.

Education Commission of the States (1983). *Calls for educational reform: A summary of major reports*. Denver, CO: ECS.

Farr, R., & Casey, R. F. (1986). *Reading: What can be measured?* Newark, DE: International Reading Association.

Fisher, C. W., Berliner, D., Filby, N., Marliave, R., Cohen, L., Dishaw, M., & Moore, J. (1978). *Teaching and learning in elementary schools: A summary of the beginning teacher evaluation study*. San Francisco, CA: Far West Regional Laboratory for Educational Research and Development.

Gullickson, A. R. (1984). Teacher perspectives of the instructional use of tests. *Journal of Educational Research*, *77*, 244-248.

Haertel, E. (in press). Student and teacher perspectives on classroom testing. Unnamed monograph on the Study of Standford Schools.

Haertel, E.,Ferrara,S., Korpi, M., & Prescott,B.(1984). Testing in secondary schools: Student perspectives. Paper presented at the Annual Meeting of the American Educational Research Association.

Haney,W. (1985). Making testing more educational. *Educational Leadership*, *43*, 4-13.

Johnston, P. (1987). Assessing the process, and the process of assessment, in the language arts. In *The Dynamics of language learning*, edited by Squire, J. R., 335-357. Urbana, IL: ERIC Clearinghouse.

Kellaghan, T., Madaus, G. F., & Airasian, P. W. (1980). *Standardized testing in elementary schools: Effects on schools, teachers, and students*. Washington, DC: National Institute of Education, Dept. of Health, Education, and Welfare.

Koretz, D. (1988). Arriving in Lake Wobegon: Are standardized tests exaggerating achievement and distorting instruction?. *American Educator*, *7*, 8-52.

Linn, R.L. (1986). Educational testing and assessment. *American Psychologist*, *41*, 1153-1160.

Linn, R. (1985). Standards and expectations: The role of testing. (Summary) *Proceedings of a National Forum on Educational Reform* (p. 88-95). New York: The College Board.

Madaus, G. F. (1981). Reaction to the 'Pittsburg Papers.' *Phi Delta Kappan*, *62*, 634-636.

Madaus, G. F. (1985). Test scores as administrative mechanisms in educational policy. *Phi Delta Kappan*, *66*, 611-618.

Ordovensky, P. (1983). Time spent in preparation for testing. *USA Today*, June 10, 1983.

Pipho, C. (1985). Tracking the reforms, Part 5: Testing--Can it measure the Success of the reform movement? *Education Week*, May 22, 1985, p.19.

Popham, W. J., Cruse, K. L., Rankin, S. C., Sandifer, P. D., & Williams, P. L. (1985). Measurement driven instruction: It's on the road. *Phi Delta Kappan*, *66*, 628-634.

Popham, W. J., & Rankin, S. C. (1981). Minimum competency tests spur instructional improvement. *Phi Delta Kappan*, *62*, 637-639.

Ruddell, R. B., & Kinzer, C. (1982). Testing preferences and competencies of field educators. In *New Inquiries in Reading Research and Instruction*. J.A. Niles and L.A. Harris (Eds.). 31st Yearbook of the National Reading Conference, Rochester, NY: National Reading Conference. pp.196-199.

Ruddell, R. B. (1985). Knowledge and attitudes toward testing: Field educators and legislators. *The Reading Teacher*, *38(6)*, 538- 542.

Salmon-Cox, L. (1981). Teachers and standardized achievement tests: What's really happening? *Phi Delta Kappan*, *62(9)*, 631-634.

Sproull, L., & Zubrow, D. (1981). Standardized testing from the administrative perspective. *Phi Delta Kappan*, *62*, 626-631.

Teale, W. (this volume)

Valencia, S., & Pearson, P. D. (1987). Reading Assessment: Time for a change. *The Reading Teacher*, *40*, 726-732.

Wallace, R. C., Jr. (1985). Redirecting a school system based on the measurement of learning. In W. W. Wellingham (Chair), *The integration of instruction and testing*. Symposium conducted at the meeting of the Educational Testing Service, New York.

Wang, M. C. (1988). The Wedding of instruction and assessment in the classroom. In *Assessment in the Service of Learning, Proceedings of the 1987 ETS Invitational Conference*. Princeton: Educational Testing Service, 63-79.

Yeh, J.P. (1980). *A reanalysis of test use data*. Test Use Project, Center for the Study of Evaluation, Report No. 143. Los Angeles: Center for the Study of Evaluation.

Chapter 6

Statewide Reading Assessment: A Survey of the States

Statewide Reading Assessment: A Survey of the States

Peter Afflerbach, University of Maryland

In this chapter, the nature of statewide reading assessment is described, using results from a survey of all statewide reading assessment programs. Assessment materials and related documents (e.g., assessment preparation materials, guides for administrators and teachers) were obtained from state education agencies and officers. The request for sample reading assessment materials was accompanied by a request for the following information: the nature of the assessment, the purpose of the assessment, the grade level(s) at which assessment is performed, the size of the assessment, and how assessment results are utilized.

Procedure

The reading assessment materials were examined and categorized. Each state's reading assessment program is described according to grade level(s) of assessment, nature of the tasks included in the assessment, assessment size, and purpose of the assessment. Additionally, special features related to a particular state's reading assessment program were considered. This information is included in the Guide to the Table.

Types of assessment

Of the 45 states which conduct statewide reading assessments, 24 use commercially produced, standardized, norm- referenced reading tests, including the Metropolitan Achievement Test, the California Achievement Test, the Iowa Test of Basic Skills, and the Comprehensive Test of Basic Skills.

Twenty-five states use multiple choice tests specially developed for the particular state. In some states, these tailor-made tests are used instead of, or in addition to, commercial, standardized, norm-referenced tests.

Six states currently use statewide, criterion-referenced reading assessments. Two states use the Degrees of Reading Power Test, which employs a cloze format.

Tasks included in assessment

The multiple-choice question format dominates the majority of statewide assessment forms. The questions deal with various tasks, ranging from parts of reading (e.g., decoding and vocabulary) at lower grade levels to text comprehension (e.g., inferential questions) at higher grade levels. The cloze reading tests required students to insert missing words in texts. Several new or revised statewide reading assessment instruments include open-ended questions.

Sample size of assessment

Thirty-eight states assess the reading ability of every student at a particular grade level. Seven states administer statewide reading assessment to a sample of students at a particular grade level.

Grade levels at which statewide reading assessment is administered

Across the states using assessments, reading is evaluated at every grade level. Reading is most frequently assessed in eighth grade, and 31 states conduct assessment in this grade. Kindergarten is the least frequently assessed grade with only four states assessing reading at this level. The frequency of reading assessment at grade levels K-12 is as follows:

Grade Level	Number of States that Assess Reading
K	4
1	12
2	13
3	28
4	19
5	14
6	24
7	15
8	31
9	15
10	18
11	17
12	8

Purpose of assessment

There is considerable variation in the stated purposes of statewide reading assessment. Thirty-four states include the determination of individual student strengths and weaknesses in reading as a purpose for assessment. Thirteen of these states indicate that one purpose of statewide reading assessment is the diagnosis of students' reading ability. Nineteen states consider a purpose of statewide assessment to be the determination of reading instructional program effectiveness.

Documenting students' minimum competencies in reading is a purpose of statewide reading assessment in 13 states, while certifying minimum competencies in reading is a stated purpose for assessment in 12 states. Five states consider dissemination of information related to students' reading achievement to be a purpose of assessment. One state explicitly acknowledges accountability as a purpose for administering statewide reading assessments. Finally, one state includes as a purpose the establishment of statewide administrative control over reading curricula.
New developments in statewide reading assessment.

In addition to the increase in the quantity of statewide reading assessments, there has been a change in the quality of certain statewide assessment efforts. Several states have implemented (or are planning to implement) new reading assessment components. The new developments are in part in response to reading researchers concerns about the validity of statewide reading assessments. In fact, several states have sought input from the reading research community while developing assessment instruments (e.g., Illinois, Michigan).

New developments in statewide reading assessment include paper and pencil assessments which measure students' prior knowledge for the content of the reading passage, assessments which allow students to make lookbacks while answering questions, and which include reading strategy questions. Such developments are included in recent assessment efforts in Illinois and Michigan. Additionally, these states have developed assessment instruments which seek information on students' knowledge about reading, as have New Jersey and Pennsylvania.

Assessments which include texts of varied length and type are planned for new reading assessment instruments in Pennsylvania and Texas. Additionally, Texas is planning to develop reading assessment tasks which more closely replicate everyday reading tasks which students face in the classroom. California is planning reading assessments with fewer multiple- choice items which will include oral and written student- constructed responses.

Several states including Illinois and Massachusetts use statewide reading assessments to collect information about students' literate behaviors outside of school. Readers' self-perceptions related to reading are also investigated in assessment items used in New Jersey and Wisconsin.

Finally, several states have integrated teacher surveys and teacher training with statewide reading assessment in an effort to better prepare teachers for classroom-based assessment of reading. Massachusetts is surveying teachers about their classroom practices, background and training, and classroom decision-making related to reading. Michigan is incorporating teacher training in informal assessment techniques to encourage increased teacher contributions to the assessment of reading ability.

Guide to the State Tables

The following provides a general description of the information found in the state reading assessment program descriptions.

Title of test or testing program

The title of the statewide reading assessment instrument is given at the beginning of each state entry. If more than one reading assessment instrument is used in a particular state, each assessment instrument is described.

Grade

Indicates the grade level(s) at which statewide assessment of reading is performed.

Tasks

Indicates the specific tasks which are included in the reading assessment instrument. Examples of these tasks include reading comprehension (e.g., main idea, detail, inferential), vocabulary, word recognition (e.g., decoding, structural analysis, use of context), study skills, and literacy skills. As indicated by the tasks included in the assessments, assessments are categorized as standardized, norm-referenced, criterion- referenced, cloze, or multiple-choice.

Assessment type

Indicates whether the reading assessment is comprehensive (administered to every student at a particular grade level in a particular state), or a sample (administered to a subpopulation of students in a particular state). Additionally, the type of sampling (e.g., matrix sampling) is noted when appropriate.

Purpose

Purpose statements were taken directly from state documents. The purposes of statewide reading assessment include the improvement of student learning, the improvement of instructional programs, the determination of minimum competency, the certification of graduation or promotion requirements, and the dissemination of school achievement to legislators, educators, and the general public.

Notes

This section includes information about particular statewide assessment programs which is not covered by the above categories. Information included in this section includes innovative assessment techniques, exemptions for particular populations of students, and pending changes in statewide assessment format.

The Statewide Assessment of Reading[1]

Alabama

Alabama Basic Competency Tests (BCT)

Grade: 3

Tasks: Word recognition (e.g., use of phonics, sight words); comprehension, reference and study skills

Grade: 6

Tasks: Word recognition; comprehension (e.g., main ideas, sequencing, inferencing, details); reference skills; literary skills

Grade: 9

Tasks: Vocabulary; Word Meaning; Knowledge of Word Parts; Comprehension; Reference Skills; Literary Skills

Assessment type: Comprehensive.

Purpose: To measure students' acquisition of basic reading skills identified as minimum for the particular grade level.

---

[1] Due to the changing nature of statewide reading assessment, some of the contents of this table subject to revision.

Alabama High School Graduation Exam (AHSGE)

Grades: 11 and 12

Tasks: Items include basic reading skills necessary for graduation; vocabulary, comprehension, reference skills, literary skills

Assessment type: Comprehensive

Purpose: To assure that persons granted an Alabama high school diploma have acquired minimum knowledge of basic reading skills.

Stanford Achievement Test

Grades: 1, 2, 4, and 5

Tasks: Reading subtests in word study skills, word reading

Grades: 7, 8, and 10

Tasks: Reading comprehension

Purpose: To gather information about a student's achievement in reading, and to allow for comparison of students with their peers on an individual school, system, state, and national basis.

Contact:
Eleanor Ann Raney, Reading Specialist
Student Instructional Services
State Department of Education
1020 Monticello Court
Montgomery, AL 36117

Alaska

Iowa Test of Basic Skills

Grade: 4, 6, and 8

Tasks: Reading skills subtests.

Assessment type: Comprehensive

Purpose: To identify strengths and weaknesses of instructional programs and individual students, and to provide appropriate instruction for students.

Contact:
Robert Silverman, Assessment Director
Educational Program Support
State of Alaska Department of Education
Goldbelt Place
801 West Tenth Street
PO Box F
Juneau, Alaska

Arizona

Arizona Pupil Achievement Testing

Grade: 1-8

Tasks: Iowa Tests of Basic Skills. Reading/reading comprehension subtests.

Grade: 9-12

Tasks: Tests of Achievement and Proficiency. Reading/reading comprehension subtests.

Assessment type: Comprehensive

Purpose: To identify student strengths and weaknesses, and to allow for comparison of local and state results with national norms.

Notes: At grades 1 and 12, testing is optional. School districts may elect to test either or both of these grade levels. There is a mandatory sampling at Grades 1 and 12 of at least 1,000 students. Reading scores are also used in the K-3 School Improvement Program to identify potential dropouts.

Contact:
Mr. Steve Stephens
State Testing Coordinator
Pupil Achievement Testing
Arizona Department of Education
1535 West Jefferson Street
Phoenix, AZ 85007

Arkansas

Metropolitan Achievement Test (MAT-6)

Grade: 4, 7, and 10

Tasks: Reading comprehension and vocabulary subtests.

Assessment type: Comprehensive

Purpose: To measure pupil performance in reading.

Arkansas Minimum Performance Tests (Criterion-reference tests)

Grade: 3

Tasks: Students are tested on word recognition, comprehension and 6 reference and study skills appropriate to their grade level.

Grade: 8

Tasks: Eighth grade students who do not pass the test in three attempts are denied promotion to ninth grade.

Assessment type: Comprehensive

Purpose: To measure pupil performance in basic subjects, provide teachers with diagnostic information, identify programmatic strengths and weaknesses, determine educational priorities, and to assess performance of schools and school districts in meeting state and district goals.

Contact:
Lynda C. White, Coordinator
Student Assessment and Curriculum
State of Arkansas Department of Education
4 State Capitol Mall
Little Rock, AR 72201-1021

California

Survey of Basic Skills

Grade: 3

Tasks: Contains 370 reading items, including word identification, vocabulary, literal and inferential comprehension, and study-locational skill items. Consists primarily of comprehension and vocabulary questions that are based on one familiar, high-interest passage of appropriate difficulty for third grade. Using the matrix-sampling method, students receive only a small portion (N = 9) of total test items.

Grade: 6

Tasks: Contains 418 reading items, including vocabulary, literal, inferential, interpretive, and critical/applicative comprehension, and study-locational skill items. Reading in the content areas includes passages drawn from literature, science, and social studies materials. Using the matrix-sampling method, students receive only a small portion of total test items.

Grade: 8

Tasks: Passages drawn from literature, science, and social studies materials. Items include vocabulary, literal, inferential, interpretive, and critical/applicative comprehension, and study-locational skills.

Grade: 12

Tasks: Includes vocabulary, literal and interpretive/critical comprehension, and study-locational skills. Eighteen different test forms are used, using a total of 131 reading items.

Assessment type: Sample. A matrix-sampling procedure is used in assessing reading statewide.

Purpose: To assess the effectiveness of school districts and schools in assisting students to master the fundamental educational skills. To provide information on programmatic strengths and weaknesses.

Notes: The State of California is embarking on a revision of all reading assessment instruments which are used in the California Assessment Program. The revised reading assessment will probably include fewer multiple-choice formats, and will incorporate oral and written student-constructed responses.

Contact:
Beth Breneman
English-language Arts Consultant
California Assessment Program
721 Capitol Mall
PO Box 944272
Sacramento, CA 94244-2720

Colorado

Colorado State Assessment Program

Reading assessment is in development. Reading ability was assessed as a part of pilot state testing program for grades 3, 6, 9, and 11 (every student) in April, 1986. Reading will be assessed as a part of an ability-and-achievement pilot program for grades 3, 6, 9, and 11 (5% sample) in April, 1987.

Grade: 4, 7, and 10

Tasks: Reading subtest of the Iowa Test of Basic Skills

Assessment type: Sample. Twenty-eight percent of all students at grade levels 4, 7, and 10 are assessed.

Purpose: To provide state-level data for the general public, legislators, and educators to improve student achievement, increase high school graduation rates, and increase school attendance rates.

Contact:
Wayne Martin
State Testing Director
Colorado State Department of Education
201 East Colfax Avenue
Denver, CO 80204

Connecticut

Connecticut Mastery Testing Program

Degrees of Reading Power, using a multiple choice cloze format, is part of the reading assessment).

Grade: 4, 6, and 8. Grades 9-12 participate in partial assessment.

Tasks: In addition to using the Degrees of Reading Power Test, assessment items will measure literal comprehension, inferential comprehension, and evaluative comprehension.

Assessment type: Comprehensive

Purpose: To improve the statewide evaluation of students reading skills. This will include early identification of students needing remedial education, continuous monitoring of students' performance, and testing of a comprehensive range of higher order reading skills.

Notes: The State of Connecticut is considering major changes in reading assessment in the next twelve months.

Contact: Peter Behuniak, Program Director
Connecticut Mastery Testing Program
Department of Education
Box 2219
Hartford, CT 06145

Delaware

Delaware Educational Assessment Program (DEAP)

Grade: 1-8

Tasks: Stanford Achievement Test (SAT)

Grade: 11

Tasks: Subtests include word reading (gr. 1), word study skills (gr. 1-3), and reading comprehension (gr. 1-8, 11).

Assessment type: Comprehensive

Purpose: To diagnose individual pupil strengths and weaknesses, place students in instructional groups or programs, identify curricular and instructional weaknesses, plan instruction, evaluate programs, and to provide guidance and counseling.

Notes: The DEAP conducted a pilot test of the Degrees of Reading Power test at the sixth grade level. Prior to making a recommendation for adoption of the DRP as part of the DEAP, thorough examination of the technical data is being made. The State of Delaware is also interested in assessing a sample of students at one grade level on a cyclical schedule. Feedback from the State Board of Education will determine the future emphases of the state assessment program.

Contact:
Kaye R. McCann, State Specialist
Educational Assessment
Research and Evaluation Division
Department of Public Instruction
Townsend Building
PO Box 1402
Dover, DE 19901

Florida

Minimum Student Performance Skills

Grade: 3, 5, 8, and 11

Tasks: Items include sight word vocabulary, word identification, literal and inferential comprehension, evaluative comprehension. Items also assess students' understanding of the purposes of reading.

Assessment type: Comprehensive

Purpose: For grades 3, 5, and 8, minimum performance skills indicate whether or not student is ready for promotion. Grade 11 minimum performance skills represent the minimum expectations for high school graduates, and must be successfully completed for high school graduation.

Notes: State of Florida standards are being revised, but have yet to be approved. No date has been set for approval.

Contact:
Lea-Ruth C. Wilkens, PhD
Reading Program Specialist
Florida Department of Education
Tallahassee, FL 32399

Georgia

Criterion Referenced Tests

Grade: PreK

Tasks: Concepts for reading

Grade: K

Tasks: Reading readiness skills

Grade: 1, 3, 6, and 8

Tasks: Literal comprehension, inferential comprehension, problem solving

Assessment type: Comprehensive

Purpose: Pre-K, K:To determine progress in reading readiness skills.
1:To evaluate student progress in reading.
3:To determine 4th grade placement; to evaluate student progress in reading.
6:To evaluate student progress in reading and mathematics.
8:Course planning for 9th grade; to identify those "at risk" in relation to high school Basic Skills Test.

Basic Skills Test

Grade: 10

Tasks: Items requiring literal comprehension, inferential comprehension, and problem solving.

Assessment type: Comprehensive

Purpose: The Basic Skills Test is considered part of the Criterion Referenced Test Program, even though it is a "minimum competency" test. The purpose is to assess minimal mastery of specific competency performance standards.

Iowa Test of Basic Skills

Grade: 2, 4, 7, and 9

Tasks: Reading subtests

Assessment type: Comprehensive

Purpose: To evaluate student progress in relation to a national norm group.

California Achievement Test (CAT), Form E, Level 10

Grade: Kindergarden

Tasks: Visual and sound recognition skills are tested.

Assessment type: Comprehensive

Purpose: Test results are used as part of the determination for readiness for promotion to first grade.

Contact:
Elizabeth Creech
Coordinator of Student Assessment
Georgia Department of Education
Office of Planning and Development
Twin Towers East
Atlanta, GA 30334

Hawaii

Stanford Achievement Test

Grade: 3, 6, 8, and 10

Tasks: Vocabulary:The student reads an incomplete statement, and from a list of four, selects a word which best completes the sentence. Reading comprehension: The student reads a passage and selects answers which best complete statements about the passage.

Spelling:The student is presented with four choices and must select the word that is spelled incorrectly.

Language:The student must complete items related to grammar, capitalization, punctuation, sentence structure, and dictionary skills.

Assessment type: Comprehensive

Purpose: To assist students, improve instruction, and upgrade programs.

Notes: Locally developed Competency-Based Measures are also used.

Contact:
Dr. Selvin Chin-Chance
Test Development Section
Office of the Superintendent
3430 Leahi Avenue
Building E, 1st Floor
Honolulu, Hawaii 96815

Idaho

Idaho Proficiency Test (IPT)

Grade:	8
Tasks:	This is an objective-referenced test which may be administered at the school or district level on a voluntary basis; reading items include following directions, using context to determine word meaning, identify sequence of events, perceive cause/effect relationships, make inferences, identify author's purpose, recognize main idea, use reference skills, make classifications and lists, interpret maps and diagrams.
Assessment type:	Comprehensive
Purpose:	To assess student mastery of those basic skills which represent essential academic prerequisites for graduation. To supply diagnostic information for use in combination with other evaluative data in adapting instructional materials and practices to accommodate individual student deficits. To provide supplemental information which may be of use in evaluating local curriculum and instructional practices, screening students for special programs, developing student schedules and making differential assignments within classes. To identify student performance trends over time. To communicate school accomplishments and continuing needs to various publics. To serve as source of information in determining State Department of Education technical assistance priorities.

Tests of Achievement and Proficiency

Grade:	11
Tasks:	Standardized, norm-referenced test. Reading comprehension subtest includes competence in reading for information from passages similar to those assigned in social studies, literature and the sciences, and materials such as labels, advertisements and newspapers, which are encountered out of school.
Assessment type:	Comprehensive

Purpose: To appraise student progress toward accomplishment of widely accepted secondary school goals in basic content areas.

Contact:

Reading Assessment Director
State Department of Education
650 West State Street
Boise, ID 83720

Illinois

Illinois Goal Assessment Program

Grade: 3, 6, 8, and 11

Tasks: Items include topic familiarity (prior knowledge) passage and constructing meaning, reading strategies, and survey of literary experience items.

Assessment type: Comprehensive

Purpose: To determine student reading ability.

Notes: Topic familiarity questions are intended to measure students' prior knowledge for the reading passage included in the assessment. Constructing meaning questions may have 1, 2, or 3 correct answers. Students receive partial credit for identifying some (1 out of 2 or 3, or 2 out of 3) of the "correct" responses.

Students are allowed to look back at the passage while answering questions. Reading strategy questions require the student to evaluate how the use of a particular reading strategy (e.g., re- reading) might help in answering questions. Literary experience questions ask students to report on their literacy activities in four areas: in-school activities, out-of-school activities, strategies used while reading and writing, and various uses of reading and writing.

Scores are not reported at the student or classroom level, and only at the school-building and district level. Students read 1 of 6 full length passages (3 narrative, 3 expository) with passages (or forms) being rotated within each class.

Contact:
Eunice Greer, Student Assessment
Illinois State Department of Education
100 North First Street
Springfield, IL 62777

Indiana

Indiana Statewide Testing for Educational Progress (ISTEP)
California Achievement Test (CAT), Form E, Basic Battery

Grade: 1

Tasks: Word analysis, vocabulary, comprehension, language expression

Grade: 2, 3, and 6

Tasks: Word analysis, vocabulary, comprehension, spelling, language mechanics, language expression

Grade: 8, 9, and 11

Tasks: Vocabulary, comprehension, spelling, language mechanics, language expression, study skills

Language Arts Supplement

Grade: 1 and 2

Tasks: Reading skills applied to everyday life--interpreting signs and symbols.

Grade: 3

Tasks: Reading skills used to gather and analyze information--locate and use parts of books; Reading skills applied to everyday life--interpreting labels.

Grade: 8 and 11

Tasks: Developing strategies for making independent evaluations of literary works--recognizing features of genres and recurring conventions of literary works.

Assessment type: Comprehensive

Purpose: To improve the educational opportunities of Indiana students. To provideIndiana schools with a means of assessing their overall educational programs in order to promote effective learning by all students. The Language Arts Supplement is included at grade levels 1, 2, 3, 8, and 11 to provide more complete coverage of the language arts proficiency statements.

Contact:
Dr. William Strange
Senior Officer
Center for School Assessment
Room 229 Statehouse
Indianapolis, IN 46204-2798

Iowa

No statewide assessment of reading. Local school districts decide which assessment measures to use, what grade levels to assess, how often and which parts of the text selected to use.

Tests often used by local school districts include: Iowa Tests of Basic Skills, Iowa Tests of Educational Development, the American College Testing Program, and the Scholastic Aptitude Test.

Contact:
Dr. Carol Alexander Phillips
Consultant, Reading
Department of Public Instruction
Grimes State Office Building
Des Moines, Iowa 50319

Kansas

Kansas Minimum Competency Testing Program

Grade: 2, 4, and 6

Tasks: Objectives are designed to address specific reading skills considered to be necessary before a student could be expected to achieve success at the next grade level. Reading skill items include word identification, use of context (grade 2); word identification, use of dictionary, identifying main idea (grade 4); identifying antonyms, identifying main idea, identifying sequence (grade 6).

Grade: 8 and 10

Tasks: Objectives are designed to address reading skills needed to function competently in adult society. Reading skill items include critically evaluating advertisements, identifying implied main idea (grade 8); following specific directions, identifying facts and opinions, determining the author's purpose (grade 10).

Assessment type: Comprehensive. Exemptions include students enrolled in special education programs which provide entirely non-academic and non-vocational activities and anyone who cannot read, understand or speak the English language.

Purpose: To provide a means of identifying students who have not attained a level of minimum competency so that remediation can be provided. Results are also designed to provide a statewide indicator of student performance on the tested competencies.

Notes: A revised program will likely be implemented in the 1990-1991 school year.

Contact: Bert Jackson
Testing Specialist
Kansas State Education Department
Kansas State Education Building
120 East Tenth Street
Topeka, KS 66612

Kentucky

Comprehensive Test of Basic Skills, 4th Edition, Benchmark Version

Grade:	K, 1, 2, 3, 5, 7, and 10
Assessment type:	Comprehensive
Purpose:	To ensure each student's right to acquire the basic knowledge and skills needed to complete high schools and enter college or the work force; to guarantee that all students had access to programs and services appropriate to their educational needs, to aid districts in developing educational improvement plans.
Notes:	Students in special education programs are exempt if recommended by student's admission or release committee. Special education students' answer documents are scored separately from those of regular students.
Contact:	Kay Vincent Reading Assessment Specialist Kentucky Department of Education Capitol Plaza Tower Frankfort, KY 40601

Louisiana

Development of new reading assessment instruments is underway. Upon entry into kindergarten for the first time, each child will be given a nationally recognized, individually administered readiness test. At the kindergarten level, the purpose will be to determine developmental readiness and to plan instruction. At grades 3, 5, and 7, all students will be administered the state-developed, criterion-referenced tests based on grade level skills.

These tests include items on vocabulary, phonetic analysis, structural analysis, comprehension, and study skills. In addition, grade 3 and grade 7 students will be administered the National Assessment of Educational Progress test for the purpose of comparing the achievements of Louisiana students with those of the nation and southern region. Three grade levels will be administered a nationally recognized norm-referenced test to compare individual student-, school-, district-, and state-level performances with a national norm.

At grade 11, all students will be administered a core curriculum test. A passing score on this test will be a requirement for high school graduation. The test will also be diagnostic and prescriptive relative to the core curriculum and serve as a basis for determining remediation needs.

Assessment type:	Comprehensive
Purpose:	To provide information about the quality of teaching and learning, about student achievement on grade-level skills. Additionally, the assessment program will provide for early identification of developmental and/or academic deficiencies of children entering school, and the proficiency of students exiting high school.
Contact:	Rebecca S. Christian Louisiana Department of Education Accountability/Assessment PO Box 94604 Baton Rouge, LA 70804-9064

Maine

Maine Educational Assessment

Grade: 4, 8, and 11

Tasks: Tasks involve items which measure performance in passage length, passage type, and reading objectives.

Assessment type: Comprehensive. Cognitive tests (objectives 1 & 2) and questionnaires (objective 3). Ten short answer questions which deal with reading. Forty common test items will be multiple-choice format; 10 will be open-ended. Matrix-sampling is used to broadly assess reading at the school level.

Purpose: To evaluate how the student comprehends what is read, manages the reading experience (objectives 1 & 2), and values reading (objective 3).

Contact:
A. Frederic Chaney
DECS
Division of Assessment
State House Station #23
Augusta, ME 04333

Maryland

California Achievement Test

Grade: 3

Tasks: Reading comprehension (literal, interpretive, and critical comprehension)

Grade: 5 and 8

Tasks: Reading comprehension

Assessment type: Sample. 60,000 students at every grade level.

Purpose: To provide local school systems with diagnostic evaluation of their instructional programs, to certify student acquisition of graduation prerequisites in reading skills, to make instructional improvements.

Maryland Functional Reading Test

Grade: 7

Tasks: Diagnostic information only. Tests five domains: Following directions, gaining information/main idea, locating information, gaining information and details, and understanding forms.

Grade: 9

Tasks: Provides diagnostic and certification information. Tests the five domains of: Following directions, gaining information and main idea, locating information, gaining information and details, and understanding forms.

Assessment type: Comprehensive

Purpose: Used for diagnostic and certification purposes. Those students that fail receive appropriate instructional assistance and may take the test twice a year, October and April, for the four years. Students must pass the test in order to receive a Maryland High School Diploma.

Contact:

John A. Johns, Program Assessment Specialist
Maryland State Department of Education
200 West Baltimore Street
Baltimore, MD 21201-2595

Massachusetts

Massachusetts Educational Assessment Program (MAEP)

Grade: 4 and 12

Tasks: Vocabulary, Inferential Comprehension (External 8 perspective; Internal perspective), Study skills

Assessment type: Comprehensive

Purpose: To identify program strengths and weaknesses.

Notes: Also included in the MAEP are a series of questionnaires for students, teachers, and principals. Student questionnaires concern attitude towards school, classroom activities, outside activities, and background. Teacher questionnaires concern background and training, school practices, decision-making, and classroom activities. Principal questionnaires concerned schoolwide variables affecting education, such as experience and longevity of the teaching staff, nature of curriculum development activities in the school and district, setting of school standards for instruction and student conduct, and availability of supplementary facilities and personnel in the school. Several questions in the principal questionnaires relate to questions in the teacher questionnaires, and can be used to provide a comparative analysis of responses.

Basic Skills Testing Program

Grade: 3, 6, and 9

Tasks: Word recognition, vocabulary, literal comprehension, inferential comprehension, critical reading skills, study skills.

Assessment type: Comprehensive

Purpose: To identify students not meeting standards in essential reading skills.

Contact: Dr. Elizabeth Badger

Bureau of Research and Assessment
The Commonwealth of Massachusetts
Department of Education
1385 Hancock Street
Quincy, MA 02169

Michigan

Michigan Educational Assessment Program (MEAP)

Grade: 4, 7, and 10

Tasks: Multiple choice and open ended questions. Test items will address following reading related matters: constructing meaning; knowledge about reading, including goals and purposes, reader/textual/contextual factors that influence reading; reading strategies; attitudes and self-perceptions related to reading; and, topic familiarity, or students' understanding of concepts key to an understanding of the passages before they read them.

Assessment type: Comprehensive

Purpose: The goal of the MEAP is to translate the research in reading underlying the new philosophy of reading and the new objectives into an assessment that is useful for instructional planning.

Notes: Reading assessment instrument is currently being revised. A new reading assessment instrument is in development. The new MEAP tests will be designed as "broad-gauged measures which reflect the goals of reading instruction as closely as possible". The majority of the assessment will use multiple- choice items. Nontraditional measures will include open-ended questions, used on a sample basis. Tests will include topic familiarity (prior knowledge) assessment questions. In addition, informal assessment techniques, for use by teachers in their classrooms, are being developed.

Contact:
Edward Roeber, Supervisor
Michigan Department of Education
Michigan Educational Assessment Program
Lansing, MI 48909

Minnesota

Grade:	4 and 8
Tasks:	Phonics, word identification using context, multiple word meanings; comprehension, including main idea, and detail. Following directions, reading graphs, alphabetization, using reference materials.
Grades:	11
Tasks:	Syllabication, comprehension, including main idea, and detail. Determining fact and opinion, following directions, reading graphs and maps, alphabetization, using reference materials.
Assessment type:	Sample
Purpose:	To analyze the curriculum, evaluate curriculum strengths and needs, and to improve student learning.
Notes:	Local school districts are given the option of using state-developed assessment materials for program development.
Contact:	Reading Assessment Specialist Minnesota Department of Education Capitol Square 550 Cedar Street St. Paul, MN 55101

Mississippi

Basic Skills Assessment Program (BSAP)

Grade:	3, 5, and 8
Tasks:	Items assess the following skills: recognizing frequently used written words; using initial sounds and context clues to predict a word in a sentence; identifying prefixes and suffixes; recognizing singular and plural forms of words; associating selected written words with the literal meanings; associating words which are same or opposite in meaning; interpreting materials; alphabetizing words using the first two letters in words; using a table of contents to locate specified information; and following written directions.
Assessment type:	Comprehensive. Criterion-referenced.
Purpose:	Comparison of schools and school districts; outcome measures for accreditation.

Functional Literacy Examination (FLE)

Grade:	11
Tasks:	Items assess the following items: associating words and phrases with their literal meanings; identifying selected written abbreviations and symbols; interpreting materials; analyzing written materials; selecting newspaper/telephone directory information; following written directions.
Assessment type:	Comprehensive. Criterion-referenced.
Purpose:	Beginning in 1990, students must pass the Functional Literacy Examination to graduate. The FLE also provides for comparison of schools and school districts, and for outcome measures for accreditation.

Stanford Achievement Tests

Grade: 1 and 4

Tasks: Reading comprehension, vocabulary, and word study skills subtests.

Grade: 6

Tasks: Reading comprehension and vocabulary subtests.

Assessment type: Comprehensive

Purpose: To plan for student remediation, to modify instructional plans, to allow for district comparison of results to the state and nation, for outcome measures for accreditation.

Contact:
Mrs. Lucy Rushing
Statewide Testing Office, Suite 805
State of Mississippi Department of Education
PO Box 771
Jackson, MS 39205

Missouri

Reading assessment instrument has been developed. Core competencies and desired learner outcomes have been developed at grade levels 2-10. Additionally, each local school district is required to use criterion referenced tests as part of the local testing program.

Assessment type: Department sample. Approximately ten percent of students will participate in the assessment. Ninety percent of all school districts will use the reading assessment instrument.

Purpose: To determine the strengths and weaknesses of students. To develop educational materials and in-service teacher training as indicated by the assessment.

Contact:
Dr. Grace McReynolds
Curriculum Consultant
Department of Elementary and Secondary Education
PO Box 480
Jefferson City, MO 65102

Montana

No statewide assessment of reading. Schools are encouraged to use standardized achievement tests, from 2nd grade through 11th grade. Schools use a variety of standardized tests, including the Iowa Tests of Basic Skills.

Contact:
Edward Eschler
English/Language Arts Specialist
Basic Instructional Services
Office of Public Instruction
State Capitol
Helena, MT 59620

Nebraska

Nebraska Assessment Battery of Essential Learning Skills (N-ABELS)

The reading component of N-ABELS is primarily concerned with decoding and is a summative instrument, requiring the student to demonstrate decoding ability by reading aloud a selection based on common vocabulary. There is no assessment of comprehension.

No standardized statewide assessment. Local districts use standardized tests for reading. The tests used vary widely, as do the conditions under which testing is conducted.

Contact:
Sharon Meyer, Elementary Consultant
Approval and Accreditation
Nebraska Department of Education
301 Centennial Mall South
Box 94987
Lincoln, NE 68509

Nevada

Comprehensive Test of Basic Skills (CTBS)

Grade:	3, 6, and 9

Nevada High School Proficiency Exam

Grade:	9 and 12
Tasks:	Items include word meaning, main idea/details, time sequence, compare/contrast, cause/effect, fact/opinion, outcome/conclusion.
Assessment type:	Comprehensive
Purpose:	To determine student strengths and weaknesses, to provide remedial help when appropriate, and to certify minimum competency for graduation.
Contact:	Dr. George Barnes Planning, Research, and Development Branch Nevada Department of Education Capitol Complex Carson City, NV 89710

New Hampshire

California Achievement Test (CAT)

Grade: 4, 8, and 10

Tasks: Reading subtests

Assessment type: Comprehensive

Purpose: To develop a statewide profile of student performance in reading. To inform the public on how well students perform certain tasks taught in schools. To provide technical assistance to school districts for instructional improvement.

Contact:
James Carr, Consultant
Donna M. Cavalieri, Curriculum Supervisor
Guidance, Testing, and Evaluation
Department of Education
State Office Park South
101 Pleasant Street
Concord, NH 03301

New Jersey

High School Proficiency Test (HSPT)

Grade: 9

Tasks: Graduation test

The High School Proficiency Test includes reading, writing, and mathematics. Reading items include literal comprehension/vocabulary skills (e.g., identify synonyms, identify details that support the main idea, identify events in sequence, identify the meaning of unfamiliar words from context), inferential comprehension/vocabulary skills (e.g., infer the main idea, draw a conclusion, distinguish between fact and opinion, synthesize information, make judgments), and study skills (e.g., locate information from a table of contents, select words from a specific dictionary page, complete an outline by selecting topic/detail.

Assessment type: Comprehensive

Purpose: To raise educational standards, improve the quality of education, and to better prepare students academically for their future.

Notes: Handicapped students participate in HSPT unless specific exemption is included in student's Individualized Education Program.

A new eleventh grade test is being developed. A draft of the basic test blueprint is as follows:

The eleventh grade HSPT for reading will present four types of texts and questions appropriate for each text type. The four types of text selected include the following: narrative, informational, persuasive/argumentative, and workplace texts.

Test questions will tap students' capacities to comprehend implicit as well as explicit meaning and to apply information from the text to new situations and contexts. In addition to multiple-choice comprehension items, the test is likely to include items which assess students' knowledge about reading and self-perceptions as readers. Concern about the issue of prior knowledge will narrow the range of topic areas to those connected directly with required course work. In addition, open-ended questions, advance organizers, and glossaries are under consideration as accompaniments to the test's texts.

Test items will be developed during 1989-1990. There will be three years of "due notice" testing. The test will be a graduation requirement for the class of 1995. An 8th grade "early warning" test is being developed concurrently. This is not intended to be a gate-keeping test, but is planned as a means for identifying students who will require remedial attention before taking the eleventh-grade test. A writing test will complement the reading test. The writing test will require students to compose and edit texts related to the reading test both in terms of text type and topic.

Contact:

Dr. Katheryn J. McGettigan
Project Coordinator
Reading Assessment/Instruction
New Jersey State Department of Education
225 West State Street
Trenton, NJ 08625-0500

New Mexico

New Mexico Reading Test

Grade:	1, 2, and 10
Tasks:	Assessment includes a sub-test from the Comprehensive Test of Basic Skills (CTBS), Form U, in addition to custom designed features for the state; comprehension is the focus of assessment at grades 1 and 2. At grade 10, applications of reading for academic, career, and personal enjoyment are assessed in a minimum competency test.
Assessment type:	Comprehensive
Purpose:	To identify student strengths and weaknesses, to provide remedial instruction, and to certify competency for graduation.
Notes:	Recently completed legislation requires that the reading assessments at grades 1 and 2 be re-examined for appropriateness and validity; changes in assessment are anticipated.
Contact:	James K. Abram, Language Arts Consultant State of New Mexico Department of Education Education Building Santa Fe, NM 87051-2786

New York

Degrees of Reading Power

Grade: 3, 6, 8, and 9

Tasks: Students read a series of passages of increasing difficulty. Seven words are omitted from each passage, and student selects most appropriate word from the five alternatives provided for each deleted word.

Assessment type: Comprehensive

Purpose: To evaluate student's current level of achievement in reading; to determine the most difficult prose text a student can profitably use in instruction and in independent reading; to measure growth in the ability to read with comprehension; to determine statewide trends in student's ability to read with comprehension; to indicate the extent of compensatory or remedial help, if any, that a student might need in order to achieve success on the Regents competency tests in reading.

Preliminary Competency Test

Grade: 3 and 9

Tasks: Students read a series of passages of increasing difficulty. Seven words are omitted from each passage, and the student selects the most appropriate word from the five alternatives provided for each deleted word.

Assessment type: Comprehensive

Purpose: To assure achievement in minimum competency required for the high school diploma.

Regents Competency Test

Grade: 11 and 12

Tasks: Students read a series of passages of approximately 300 words each. Seven words are omitted from each passage, and student selects most appropriate word from the five alternatives provided for each deleted word.

Assessment type: Comprehensive

Purpose: To assure achievement of minimum competency required for high school diploma.

Contact:
Carolyn Byrne, Director
Division of Educational Testing
State Education Department
Albany, NY 12234

North Carolina

North Carolina Annual Testing Program
California Achievement Test, 1985 Edition

Grade: 1, 2, 3, 6, and 8

Tasks: Reading and language arts subtests.

Assessment type: Comprehensive

Purpose: To help local school systems and teachers identify and correct student needs in basic skills. The results of the CAT are also used to identify students whose CAT results ranked in the bottom quarter nationally for their grade level, and who were unlikely to succeed at the next grade level without remediation. These students are then administered the North Carolina Minimum Skills Diagnostic Tests.

North Carolina Minimum Skills Diagnostic Tests

Grade: 3, 6, and 8

Tasks: Reading items measuring most basic skills needed for the next grade.

Assessment type: Administered to all students falling below the twenty-fifth percentile of the California Achievement Test.

Purpose: To measure minimum skills required for success at next grade level.

North Carolina Competency Test

Grade: 10

Tasks: Minimum competency reading tasks

Assessment type: Comprehensive

Purpose: High school graduation requirement.

Contact:
Reading Assessment Specialist
State of North Carolina
Superintendent of Public Instruction
Raleigh, NC 27611

North Dakota

No statewide assessment of reading. Most schools incorporate standardized tests into the curriculum, including the Iowa Test of Basic Skills, or the SRA Survey of Basic Skills.

Contact:
Pat Herbel, Director
Elementary Education
Department of Public Instruction
State Capitol
Bismarck, ND 58505

Ohio

Each school district has a required competency-based education program in reading, and English/language arts. Each district decides which objectives in its locally developed reading program are to be emphasized. These objectives are written in behavioral terms and are assessed throughout the year by classroom teachers, who also provide intervention as needed. Each district is required to test students a minimum of three times during their school career in reading: once in grades 1-4, once in grades 5-8, and once in grades 9-11. The test results are used by the district and do not have to be reported to the Ohio State Department of Education.

Notes: Ohio will implement additional statewide assessment of reading in 1989 and 1990. Two new statewide testing programs will be implemented. In the first, Ohio school districts will measure each student's reading in terms of the student's ability. Students in grades four six, and eight will be assessed in reading. The state board of education has adopted a list of standardized tests from which all districts must choose. Results must be sent to the Ohio Department of Education. The Department of Education will aggregate and report test results by grade and test area. To earn a diploma, students will demonstrate at least a ninth grade proficiency level in reading. By demonstrating at least a twelfth grade proficiency level in reading, the student may earn one of several types of diplomas. The state will provide for the test, its scoring, and reporting results to the schools.

Contact:
Susan E Gardner, Consultant
Department of Education
Columbus, OH 43215

Oklahoma

Metropolitan Achievement Test/6e

Grade: 3

Tasks: Vocabulary, word recognition, comprehension

Grade: 7 and 10

Tasks: Vocabulary, comprehension

Assessment type: Comprehensive

Purpose: To afford a component for use along with other pertinent data in evaluating the effectiveness of the public schools as shown by the competence and progress of pupils in basic skills.

Notes: Legislation may expand the grade levels at which the MAT/6E is administered from 3, 7, and 10 to 3, 5, 7, 9, and 11. Tests would continue to be administered in census fashion. In any event, the state will continue to use norm-referenced assessments.

Contact:
Reading Assessment Director
Oklahoma State Department of Education
Oliver Hodge Building
2500 North Lincoln Blvd.
Oklahoma City, OK 73105-4599

Oregon

Oregon Statewide Assessment

Grade:	8
Tasks:	Items include word meaning, main ideas, supporting details, facts and opinions, use of instructional materials, inferential comprehension, inferencing, evaluation of written material.
Assessment type:	Sample. Approximately 15% of schools will be assessed.
Purpose:	To provide state level information on the status of student achievement in the state and feedback to state specialists and committees on the achievement of students related to state curriculum goals in order to set state priorities for improvement.
Contact:	Office of Policy and Program Development Oregon Department of Education 700 Pringle Parkway SE Salem, OR 97310-0290

Pennsylvania

Testing of Essential Learning Skills (TELS)

Grade:	3, 5, and 8
Tasks:	Vocabulary, literal comprehension, inferential comprehension life/study and reference (for all three grade levels).
Assessment type:	Comprehensive
Purpose:	To provide early identification of those students who need remedial instruction.
Notes:	The TELS will not be used after 1989. Beginning in 1990, the reading test will be based on the following definition of reading: a dynamic process in which the reader interacts with the text to construct meaning. Inherent in constructing meaning is the reader's ability to activate prior knowledge, use reading strategies and adapt to the reading situation. Based on this definition, items in the forthcoming reading assessment will examine prior knowledge, ability to construct meaning/comprehend text, ability to use reading strategies, reading habits and attitudes. Narrative and informational passages to be included in the assessment will be age-, interest-, and readability- appropriate for the grade level being tested. The narrative passages will be complete works of varying length, from a maximum length of 500 words at grade 2, to a maximum length of 2000 words at grades 9 and 10. The items used with the informational passage will be at three levels of processing: explicit, which require the student to identify, locate, or confirm information directly stated in the passage; implicit, which require the student to use textual information and prior knowledge to construct meaning and make inferences; and, extended, which require the student to respond to and think beyond the text.
Contact:	Leann Miller Educational Assessment Specialist Division of Educational Testing and Evaluation Bureau of Educational Planning and Testing Pennsylvania Department of Education 333 Market Street Harrisburg, PA 17126-0333

Rhode Island

Rhode Island State Assessment Program
Metropolitan Achievement Test

Grade: 3

Tasks: Word recognition, vocabulary and reading comprehension subtests.

Grades: 6, 8, and 10

Tasks: Vocabulary and reading comprehension subtests.

Assessment type: Comprehensive

Purpose: To provide data for both educational program decisions that will directly benefit individual students as well as data to guide the development of educational policy and curriculum.

Contact:
Dr. Pasquale DeVito
Evaluation and Testing
Department of Elementary and Secondary Testing
State of Rhode Island
22 Hayes Street
Providence, RI 02908

South Carolina

Basic Skills Assessment Program (BSAP)

Grade: 1, 2, 3, 6, and 8

Tasks: Criterion referenced assessment reading items.

Assessment type: Comprehensive

Purpose: To determine student achievement, strengths and weaknesses, and to improve instruction.

Exit Exam

Grade: 10

Tasks: Basic competency skills in reading.

Assessment type: Comprehensive

Purpose: Requirement for high school graduation.

Contact:

Reading Assessment Specialist
Student Assessment Unit
Department of Education
Columbia, SC 29201

South Dakota

No statewide assessment of reading.

Contact:

Reading Assessment Specialist
Richard Kniep Building
700 North Illinois Street
Pierre, SD 57501-2293

Tennessee

Basic Skills First Achievement Test

Grade: 3, 6, and 8

Tasks: Criterion-referenced items in reading comprehension.

Assessment type: Comprehensive

Purpose: To modify instructional programs and identify student strengths and weaknesses.

Stanford Achievement Tests

Grade: 2, 5, 7, 9, and 12

Tasks: Reading subtests.

Assessment type: Comprehensive

Purpose: To modify instructional programs and identify student strengths and weaknesses.

Tennessee Proficiency Test

Grade: 9

Tasks: Language arts test.

Assessment type: Comprehensive

Purpose: To assess minimum competency achievement.

Notes: As of 1990, reading will be assessed in grades 2-8 using customized tests which will yield both norm-referenced and criterion-referenced scores from each grade level test form.

Contact:

Angelia Golden
Director of State Testing
Tennessee State Department of Education
1150 Menzler Road
Nashville, TN 37210

Texas

Texas Educational Assessment of Minimum Skills (TEAMS)

Grade: 1, 3, 5, 7, 9, and 11

Tasks: Includes reading vocabulary lists, developed from five basal reading series on the state adopted list, the Dolch Basic Word List, and words with which children were most likely to be familiar.

Objective performance data and total test mastery information are reported for each student, school, district, region of the state, and the state as a whole.

Assessment type: Comprehensive

Purpose: To measure minimum competencies.

Notes: Beginning in October, 1990, Texas will be instituting a new assessment program, as yet untitled. The test will assess academic skills, rather than minimum skills. The reading portion will include a variety of text types: narrative, informative, and functional. In addition, the passages will be notably longer, from a 300 word maximum on the third-grade level to a 1000 word maximum on the exit level (grades 11 and 12). The intent is to place each tested task in a meaningful reading context which closely replicates the tasks students are being asked to do on an every-day basis.

Contact:
Patricia Sachse Porter
Director of Programs
Division of Student Assessment
Texas Education Agency
1701 North Congress Avenue
Austin, TX 78701

Utah

Comprehensive Test of Basic Skills
CTBS (Form U, Levels G and J)

Grade: 5 and 11

Tasks: Vocabulary (same-meaning words, unfamiliar words in context, multi-meaning words, missing words in context, meaning of affixes). Comprehension (passage details, character analysis, main idea, generalization, written forms, writing techniques). Different level of CTBS, including same subtests and components.

Assessment type: Sample. Fifth grade-4500 students, eleventh grade-2600 students.

Purpose: To monitor public school performance, to indicate how certain demographic factors influence the way students achieve, to provide insight into instructional approaches which result in high achievement of students.

Notes: In addition to Utah's Statewide Assessment Program, which uses a norm-referenced test to measure reading, the state has an extensive program of criterion-referenced testing in reading. In reading, end-of-level tests are available for use by Utah school districts for grades 1 through 6. Virtually all Utah school districts are using this voluntary assessment program.

Contact:
Nancy Livingston
Curriculum and Instruction/Reading
Utah State Office of Education
250 East 500 South
Salt Lake City, UT 84111

Vermont

No current statewide assessment of reading.

Notes: Statewide assessment of reading may be planned for the early or mid-1990's, following development of assessment in other areas, including writing.

Contact: Susan Carey Biggad
Elementary Reading/Language Arts Consultant
Department of Education
State Office Building
Montpelier, VT 05602

Virginia

Literacy Test (Degrees of Reading Power)

Grade: 6

Tasks: Comprehension

Assessment type: Comprehensive

Purpose: Assessment of students' ability to comprehend non- fiction reading passages.

Standards of Learning Program

Grade: Every grade level.

Tasks: Objectives reflect competencies at each specific grade level.

Assessment type: Comprehensive

Purpose: Establishes a framework for instruction and assessment by stating in objective format the reading skills and knowledge that students are expected to acquire at each grade.

Iowa Tests of Basic Skills (ITBS)

Grade: 4, 8, and 11

Tasks: Reading related skills, including comprehension. Tests of Achievement and Proficiency

Assessment type: Comprehensive

Purpose: To provide norm-referenced information on student achievement in reading.

Contact: Dr. Lois Rubin

Director of Research and Testing
Department of Education
PO Box 6Q
Richmond, VA 23216-2060

Washington

Metropolitan Achievement Test

Grade: 4, 8, and 10

Tasks: Reading and language arts subtests.

Assessment type: Comprehensive

Purpose: To establish the reading level of all fourth, eighth, and tenth graders. Assessment results may be used as an initial screening for remediation and/or gifted programs. The main purpose of the test is to give a sense of how students are doing statewide.

Contact: Fred Bannister, Supervisor
Reading/Language Arts
Old Capitol Building, FG-11
Olympia, WA 98504

West Virginia

Cognitive Abilities Test/3e

Grade: 3 and 9

Tasks: Reading subtests.

Comprehensive Test of Basic Skills (CTBS)

Grade: 3, 6, 9, and 11

Tasks: Vocabulary and comprehension subtests.

Assessment type: Comprehensive

Purpose: To provide information to students, parents and educators that assists in the decision-making process related to educational and career planning; for the evaluation, planning and improvement of educational programs.

Notes: Beginning in 1990-1991, the State of West Virginia will change to a series of criterion-referenced tests, administered in grades one through eight.

Contact:

Larry C. Gabbert
Coordinator, State-County Testing Programs
Capitol Complex, Building 6
Department of Education
Charleston, WV 25305

Wisconsin

Third Grade Reading Test

By state legislative mandate, the Wisconsin Department of Public Instruction is required to develop a third grade reading assessment test which was administered for the first time during the 1988-89 school year, and will be annually. The test was taken by all third graders in the state. Comprehension scores will be reported in relation to a statewide performance standard. The intent of the third grade assessment is early identification of students in need of remediation. The test is scheduled to be administered for the first time in April, 1989, to approximately 59,000 students. It is designed to gather five types of information: general reading behaviors, reading strategies, prior knowledge, comprehension, and passage-specific attitudes and self-perceptions.

Contact: Vicki Frederick, Education Specialist
Bureau for Achievement Testing
Department of Public Instruction
125 South Webster Street
Box 7841
Madison, WI 53707

Wyoming

In 1988, 20% of all 4th, 8th, and 12th grades took part in a concurrent assessment with the National Assessment of Educational Progress in reading.

Assessment type: Sample

Purpose: To determine sample student performance and program effectiveness.

Contact: Jim Lendino
Education Program Planning
Evaluation Specialist
Department of Education
Hathaway Building
Cheyenne, WY 82002

Index

Accountability 1, 8, 9, 32, 48, 75, 102, 125
Albuquerque Public Schools 44
Alternative assessment 1, 2, 42, 49, 57, 63, 68, 69, 78, 84, 93
Association for Childhood Education 36
Association for Supervision and Curriculum Development 36

Becoming a Nation of Readers 13, 60, 69, 95
Bias review 25, 26

Competencies 18, 21, 75, 97, 102, 123, 135, 155, 158
Conceptions 3, 40, 67
Connecticut 65, 69, 71, 112
Content 2, 3, 8, 12, 18-22, 24-29, 31, 36, 38, 39, 41, 42, 44, 46, 57, 65, 84-87, 93, 103, 110, 119
Criterion-referenced 2, 17-19, 28, 30, 35, 57, 79, 101, 125, 133, 153, 154, 156, 160
Curriculum 1, 2, 7-13, 17-21, 24, 30, 31, 35-38, 43, 50, 57, 59, 60, 66, 75-77, 80-82, 85-87, 90-95, 109, 118, 125, 129, 132, 135, 138, 145, 148, 150, 156
Cut scores 18

Definition 19, 57, 149
Diagnostic testing 8, 12, 18, 28-30, 49, 70, 87, 92, 94, 109, 118, 125, 127, 144
Diagnostic Workshops 28
  uses of 1, 3, 7, 8, 12, 41, 80, 90, 91, 120

Effects of testing 1, 7, 42, 79, 97
Emergent literacy 2, 39, 40, 44, 45, 52

Field-testing 24, 25
Format 2, 12, 25, 38, 42-44, 76, 78, 101, 104, 112, 126, 158

Gathering data 46, 61, 62
Georgia 2, 17-19, 21, 25, 30-32, 115, 116
Goals 2, 3, 12, 17, 36, 37, 57, 59, 60, 66-69, 77, 85, 86, 89, 92, 94, 109, 119, 131, 148

Illinois 57, 70, 75, 103, 120, 152
Indicators 9, 10, 19, 20, 70, 87

Interactive processes 9
International Reading Association 36, 50-52, 71, 72, 94, 96
Item
  development 21
  pool 18
  review 24, 26

Large-scale 12, 17, 32, 64, 75
Levels 3, 8, 18, 19, 24, 25, 62, 76, 79, 80, 84, 88, 101, 102, 108, 111, 122, 125, 135, 147, 149, 156
Literacy 2, 11, 13, 36, 38-45, 47-52, 55, 57-62, 65, 67-70, 72, 104, 120, 133, 158

Massachusetts 64, 71, 103, 129, 130
Michigan 52, 57, 103, 131
Multiple-choice 22, 24, 32, 37, 42, 49, 57, 59, 63, 69, 101, 104, 111, 126, 139

National Association of Elementary School Principals 36
National Council of Teachers of English 36
National survey of test data
  analysis of 46, 63, 78, 129
  discussion about 64
  purpose of 1, 3, 10, 23, 29, 31, 42, 101, 102, 125, 159
  results of 3, 30, 84, 87, 89, 144
New Mexico 44, 141
New York 13, 49, 51, 52, 58, 70-72, 97, 98, 142
North Carolina 35, 41, 46, 50, 144, 145

Observational data 62
Operational assessment forms 2, 26
  constructing 2, 18, 24, 41, 120, 131, 149
Oral expression 62

Pass/fail 18, 30
Pedagogy 61
Portfolios 3, 57, 58, 61, 63, 66
Prior knowledge 10, 57, 65, 103, 120, 131, 139, 149, 161
Problems 22, 39, 43, 57, 94

Project zero 66
Purposeful reader 10

Questioning 3, 10, 46, 61, 62, 69

Rasch model 29
Reporting 18, 36, 79, 82, 87, 88, 146

Sampling 21, 25, 61, 63, 67-69, 72, 78, 79, 104, 108, 110, 126
Scale 8, 12, 17, 18, 28-30, 32, 45-47, 64, 75
Specification 18
Standard setting 30, 31
Standardized tests 7, 9, 10, 36, 39, 43, 47, 48, 61, 63, 69, 76, 82, 87, 90, 92-95, 97, 135, 136, 145, 146
Standards 1, 2, 8, 18, 19, 30, 31, 50, 97, 114, 115, 129, 139, 158
Statewide reading assessment 2, 1-3, 7-9, 11-14, 17, 38, 40, 42, 46, 53, 72, 95, 99, 101-105
  concerns with 2, 9
  factors contributing to the use of 7
Statewide reading assessment programs 13, 40, 42, 46, 101
  beginning 18, 19, 25, 29, 31, 35, 37, 49, 63, 64, 68, 77, 96, 103, 133, 149, 155, 160
  developing 1, 2, 15, 17, 19, 21, 38, 103, 118, 121, 124
Strategies 10, 11, 20, 23, 35, 38, 41, 46, 47, 57, 58, 62, 64, 72, 84-86, 92, 93, 120, 121, 131, 149, 161

Teacher involvement 2
Teacher-based assessment 2, 3, 58, 60, 61, 66, 67, 69
Teacher-based information in literacy learning
  next steps for 67
  reasons for 1, 58
  role of 2, 55, 57, 60, 68, 97
Test
  construction 29
  development 18, 19, 25, 31, 117
  equating 27, 29, 30
  influence 10, 11, 50, 70, 72

Validity 1, 7, 9, 10, 12, 13, 20, 21, 43, 47, 88, 95, 102, 141
Vermont 57, 69, 70, 157

Workshop 19-31
Writer training 21-26

Young children 2, 35-45, 47-49, 51, 52